Next-Generation Statistical Genetics: Modeling, Penalization, and Optimization in High-Dimensional Data

Kenneth Lange,[1,2,3] Jeanette C. Papp,[2] Janet S. Sinsheimer,[1,2,3,4] and Eric M. Sobel[2]

[1]Department of Biomathematics, [2]Department of Human Genetics, [3]Department of Statistics, and [4]Department of Biostatistics, University of California, Los Angeles, California 90095; email: klange@ucla.edu, jcpapp@ucla.edu, janets@mednet.ucla.edu, esobel@ucla.edu

Annu. Rev. Stat. Appl. 2014. 1:279–300

First published online as a Review in Advance on November 8, 2013

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

This article's doi: 10.1146/annurev-statistics-022513-115638

Keywords

computational statistics, data mining, DNA sequence analysis, gene mapping, pedigrees

Abstract

Statistical genetics is undergoing the same transition to big data that all branches of applied statistics are experiencing, and this transition is only accelerating with the advent of inexpensive DNA sequencing technology. This brief review highlights some modern techniques with recent successes in statistical genetics. These include (*a*) Lasso penalized regression for association mapping, (*b*) ethnic admixture estimation, (*c*) matrix completion for genotype and sequence imputation, (*d*) the fused Lasso for discovery of copy number variation, (*e*) haplotyping, (*f*) relatedness estimation, (*g*) variance components models, and (*h*) rare variant testing. For more than a century, genetics has been both a driver and beneficiary of statistical theory and practice. This symbiotic relationship will persist for the foreseeable future.

1. INTRODUCTION

Gene: an inherited DNA segment whose expression ultimately leads to an observable trait

Genome: the full set of genetic material of an individual

Locus: the position of a specific DNA segment on a chromosome

Allele: one of the possible states that a gene can take; also called a variant

Variant: an allele at a locus; can be common or rare

Single-nucleotide polymorphism (SNP): a locus, usually biallelic and never more than tetraallelic, characterized by variation at a single base position

Genetics and statistics have coevolved for more than a century (Bodmer 2010), and modern statisticians are heavily involved in current genetic studies (Mechanic et al. 2012). This symbiotic relationship will likely continue well into the twenty-first century. The current review provides a limited survey of statistical genetics and its future over the next decade. The repeated revolutions in both genetics and statistics suggest that considering a longer time horizon would be foolish. Our review focuses on linkage and association studies (gene mapping) for disease traits, as well as how these studies are impacted by cheap DNA sequencing, fast parallel computing, and recent advances in statistics. It is also worth emphasizing the positive benefits genetics exerts on statistics by suggesting new problems and serving as a testing ground for new techniques. Indeed, statistical geneticists have helped usher in a new era of theoretical statistics dominated by big data and high-dimensional problems. To assist readers new to the field of genetics, we include definitions for key genetic terms and acronyms. These readers may also consult reference texts (Laird & Lange 2011, Strachan & Read 2011, Thomas 2004, Ziegler et al. 2010) for fuller explanations of genetic concepts.

The advent of high-throughput single-nucleotide polymorphism (SNP) genotyping focused statisticians' attention on several challenges: (*a*) less stringent *p*-value adjustments for multiple testing, (*b*) model selection with a wild excess of predictors over outcomes, (*c*) quality control of massive data sets, (*d*) adjustment for potential confounders such as population substructure, and (*e*) ultrafast computation of test statistics (Cantor et al. 2010). However, once statisticians successfully overcame these hurdles, along came next-generation sequencing (see sidebar, The Accelerating Pace of Genetic Data Acquisition). Although this new technology produces orders of magnitude more data, it is more error prone and creates enormous problems of data storage and manipulation (Mechanic et al. 2012). In compensation, inexpensive sequencing technology has fostered the study of rare variants (Bodmer & Bonilla 2008). Previous genomewide association studies (GWAS) dealt exclusively with common variants (see sidebar, Current Standard Genetic Data Analysis Procedures). The outcomes have been spectacular by many metrics (Hindorff et al. 2009). Unfortunately, their utility has reached the point of diminishing returns (Ku et al. 2010, Visscher et al. 2012). Although more associated SNPs are being discovered via GWAS, their

THE ACCELERATING PACE OF GENETIC DATA ACQUISITION

A dramatic scale-up of genetic data production has occurred in the last decade. Genotyping scans have increased from hundreds of markers to millions of SNPs, with parallel advances in sequencing technology. After roughly 13 years and at a cost of $3 billion, the first human genome sequence was completed in 2003. The NIH then called for technologies capable of sequencing a human genome for $1,000 or less, in a day or less. The goal is to make genome sequencing a common research and diagnostic tool. The "next-generation" sequencers increased throughput and lowered cost via reaction miniaturization and parallelization. Today we are entering a third generation of sequencing technology characterized by innovations such as single-molecule sequencing, longer reads, and new detection methods including ion semiconductors, mass spectrometry, and electron microscopy. Currently, sequencing a complete human genome costs several thousands of dollars, so labs often sequence only exomes (roughly the 1% of the genome that encodes genes) or other regions of interest. However, continuing advances in sequencing technology suggest that $1,000 or even $100 whole-genome sequences produced within hours will be available soon. It will then be routine for researchers and clinicians to obtain each subject's full 6-billion-base-pair diploid sequence. The statistical challenges of coping with this level of data will be enormous. For a more detailed overview, follow the **Supplemental Material** link from the Annual Reviews home page at **http://www.annualreviews.org**.

Supplemental Material

CURRENT STANDARD GENETIC DATA ANALYSIS PROCEDURES

To map genes that influence traits, one searches for correlations between the phenotypes and genotypes of study subjects. Genotypes may be assayed from a predefined set of candidate gene regions (hypothesis-driven) or from throughout the genome (hypothesis-free). Inexpensive, high-density, whole-genome SNP genotyping became available around the turn of the millennium. This technology allowed hypothesis-free genomewide association studies (GWAS) of unrelated individuals to become the standard gene-mapping procedure, as Risch & Merikangas (1996) predicted. DNA is extracted from each individual and interrogated via a SNP chip that yields his or her genotypes. Fortunately, competition among several vendors rapidly brought the per-chip cost down to a few hundred dollars and the per-chip yield up to several million SNPs. Genotyping error and missing data rates are typically low, considerably below 1%. In a standard GWAS, each SNP is analyzed separately for correlation with a given trait using either linear regression (for quantitative traits) or logistic regression (for qualitative traits). Today, routine GWAS can involve a few thousand unrelated individuals, each genotyped at a few million SNPs. GWAS have identified thousands of previously unknown genes for hundreds of common human diseases (Hindorff et al. 2009). Despite their many successes, standard GWAS have some inherent design limitations. For instance, predictors greatly outnumber observations, mandating stringent *p*-values; causative rare variants are therefore hard to find; and found variants are thus fairly common, implying that they have small effect sizes. In addition, affected pedigrees, which are more likely to include genes of relatively strong effect, are not analyzed in standard GWAS. Finally, interaction analyses quickly become intractable.

effects are minuscule at the population level. This state of affairs makes eminent sense because alleles with major deleterious effects are quickly eliminated by natural selection and therefore are unlikely to become common. In contrast, rare variants, with population frequencies below 1%, may have quite large effects while scattered across a single gene (Gibson 2012).

The problem with rare variants is simply their rarity. Even in large studies, many rare variants are observed in only a handful of people. This makes a simple marginal analysis of each variant impractical. However, several possible remedies exist. First, one can turn to pedigrees in which rare disease variants tend to cluster. To some degree, studies in population isolates, in which almost everyone is related, fall into this category. The earlier paradigm of linkage analysis explicitly exploits such pedigree data (Lange 2002). Second, one can combine rare variants in statistical analysis by collapsing the variants (Li & Leal 2008) or by aggregating them empirically through adaptive weights or group penalties. Here the Lasso and Euclidean penalties are useful tools (Zhou et al. 2010, 2011b). Bayesian approaches that incorporate prior biological knowledge are attractive in principle, but Markov chain Monte Carlo (MCMC) algorithms can quickly grind to a halt under the sheer mass of sequence data. Third, one can use meta-analysis, which has become a standard tool in statistical genetics because it borrows strength across studies (Cantor et al. 2010). Sequence data both increase the need for meta-analysis and the challenges in employing it (Asimit et al. 2012a, Derkach et al. 2013, Singh et al. 2013).

The two principal levers of inference in genetic analysis are relatedness and linkage disequilibrium (LD). Relatedness is problematic because paternity records, and occasionally even maternity records, are unreliable. In population isolates, extended genealogies are suspect due to cryptic relationships, particularly among pedigree founders. With dense genotyping, it is now possible to estimate kinship coefficients quickly and accurately. These empirical estimates serve as surrogates for fully traced pedigrees. Relatedness can also be assessed at the local level along the genome. In reasonably short genomic intervals, one can determine highly specific dense haplotypes. Because two individuals with the same haplotype probably inherited that haplotype from a common

Genotype: the two alleles found at a locus in an individual

Candidate gene: a gene (or SNP) that is singled out for study because of prior biological evidence or statistical analyses

Linkage: the tendency of nearby loci to be inherited together

Genetic association: the tendency of a particular genotype to be seen in one group (e.g., cases) over another group (e.g., controls)

ancestor, detailed haplotyping permits statisticians to combine the strengths of association and linkage analysis.

The questions we cover in this review include the following: (*a*) How can one perform model selection in association studies? (*b*) How can one capitalize on pedigree data in association analysis? (*c*) How can one best impute SNP genotypes in low-coverage sequencing data? (*d*) How can one estimate relatedness locally and globally along the genome? (*e*) What are good ways to include rare variants in association analyses? (*f*) What role does data mining play in genomics? Many of the methods described in this review have been or will be implemented in MENDEL, our freely available statistical genetics package (Lange et al. 2013). To keep within reasonable page limits, we largely ignore the important topics of gene expression and epigenetics. Instead, we emphasize the interaction between models and computation. Algorithms are the glue that binds these two vital ingredients. Unfortunately, not all algorithms are created equal. The best algorithms combine speed, scalability, parsimony, statistical power, and fidelity to reality. Genetics offers a concrete setting in which these tensions in computational statistics play out with maximal impact on society.

Linkage equilibrium (LE): describes loci for which the frequency of each haplotype is the product of its component allele frequencies

Linkage disequilibrium (LD): describes loci that are not in linkage equilibrium

Haplotype: a combined set of alleles from closely linked loci usually passed intact from parent to child

Cryptic relationships: unstated relationships within a sample of individuals

Coverage: number of redundant fragments read to determine a short region's sequence; multiple fragments are required due to high error rates

2. BLOCK DESCENT AND BLOCK ASCENT

A brief review cannot do justice to the scope of statistics and its relationship to genetics. Instead, we focus on a few current vignettes that illustrate our main themes of modeling, penalization, and optimization in high-dimensional data. We highlight two pillars of modern computational statistics: block relaxation in this section (de Leeuw 1994) and the MM algorithm in Section 3 (Hunter & Lange 2004, Lange 2010, Wu & Lange 2010). These two algorithmic principles guarantee that the objective function steadily ascends in maximization and steadily descends in minimization. These principles should not be viewed as competitors. In fact, they can be mixed and matched in creative ways in the same problem.

The twin notions of block descent and block ascent are conveyed by the generic term block relaxation (de Leeuw 1994). Block relaxation is a good option in high dimensions where Newton's method and Fisher scoring hit insurmountable barriers. Block relaxation divides the parameters into disjoint blocks and then cycles through the blocks, updating only those parameters within a given block at each stage of a cycle. When each block consists of a single parameter, block relaxation is called cyclic coordinate descent or cyclic coordinate ascent. Block relaxation is best suited to unconstrained problems in which the domain of the objective function reduces to a Cartesian product of the subdomains associated with the different blocks. Obviously, exact block updates are a huge advantage. Equality constraints usually present insuperable barriers to cyclic coordinate descent and ascent because parameters get locked into place. In some problems overlapping blocks are advantageous.

2.1. Lasso Penalized Regression and Association Mapping

The regression problems generated from big data often entail a vast excess of predictors over cases. This obstacle has spurred innovation in model selection and fast computation because classical methods of regression are ill equipped in this realm. One of the most profound discoveries of modern computational statistics has been the therapeutic effects of Lasso (ℓ_1) penalties (Chen et al. 1998, Claerbout & Muir 1973, Santosa & Symes 1986, Taylor et al. 1979, Tibshirani 1996). The imposition of Lasso penalties makes continuous model selection possible and avoids the computational bottlenecks of classical selection by regression. Lasso penalties force most regression coefficients to equal 0. Cyclic coordinate descent is perfectly matched to the needs of Lasso penalized regression if just a handful of predictors are ultimately selected (Friedman et al. 2007,

Wu et al. 2009, Wu & Lange 2008). The amount of computation needed to update a regression coefficient in each stage of a cycle is light, particularly if the coefficient starts and remains at 0.

Lasso penalized regression can be phrased as minimization of the objective function

$$f(\theta) = g(\theta) + \rho \sum_{j=1}^{p} |\beta_j|, \qquad 1.$$

where the loss function in ordinary regression is

$$g(\theta) = \sum_{i=1}^{m} \left(y_i - \mu - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2.$$

Here $\theta = (\mu, \beta_1, \ldots, \beta_p)^t$ is the parameter vector and $\mathbf{X} = (x_{ij})$ is the $m \times p$ design matrix. In generalized linear models such as logistic regression, the loss function becomes the negative log likelihood (Friedman et al. 2010, Wu et al. 2009). The strength of the Lasso penalty in the criterion given by Equation 1 is determined by the positive tuning constant ρ. The Lasso penalty shrinks each β_j toward the origin, discouraging models with large numbers of marginally relevant predictors. No penalty is imposed on the intercept μ because it should appear in any plausible model.

In practice, minimization of the loss function drives regression. Standard methods of ℓ_2 regression require matrix inversion, matrix diagonalization, or the solution of large systems of linear equations. These tasks take $O(p^3)$ arithmetic operations and are intractable for problems with tens of thousands of predictors. Furthermore, when the number of predictors exceeds the number of cases, the familiar $\mathbf{X}^t\mathbf{X}$ matrix is singular. Coordinate descent avoids these thorny issues and enjoys the desirable properties of simplicity, speed, and stability (Alexander & Lange 2011b, Friedman et al. 2007, Wu & Lange 2008, Wu et al. 2009, Zhou et al. 2010). The tuning constant ρ can be chosen by bracketing and by golden section search of an appropriate cross validation criterion.

The nondifferentiability of the Lasso penalty is the primary barrier to cyclic coordinate descent. This obstacle is overcome by considering the two domains $\beta_j \geq 0$ and $\beta_j \leq 0$ separately when updating β_j. Ordinarily, the objective function is convex, and the signs of its forward and backward directional derivatives at the origin determine the pertinent domain. The two derivatives along the jth coordinate direction $\mathbf{e}_j$ amount to

$$\mathrm{d}_{\mathbf{e}_j} f(\theta) = \frac{\partial}{\partial \beta_j} g(\theta) + \rho \begin{cases} 1 & \beta_j \geq 0 \\ -1 & \beta_j < 0 \end{cases}$$

for the forward direction and

$$\mathrm{d}_{-\mathbf{e}_j} f(\theta) = -\frac{\partial}{\partial \beta_j} g(\theta) + \rho \begin{cases} -1 & \beta_j > 0 \\ 1 & \beta_j \leq 0 \end{cases}$$

for the backward direction. If either directional derivative is negative, then one solves for the minimum in that direction. Otherwise, β_j is set to 0.

In many settings, it is reasonable to couple regression coefficients so they enter a model as a group. For instance, in GWAS, one might want to group the SNPs within a gene or the genes within a biochemical pathway. One can coordinate the selection of predictors by adding group penalties that preserve the convexity of the objective function and retain consistency with cyclic coordinate descent (Friedman et al. 2010, Meier et al. 2008, Yuan & Lin 2006, Zhou et al. 2010). The conceptually simplest way to group regression coefficients is to add Euclidean distance penalties. Suppose that p predictors are partitioned into a collection G of nonoverlapping but exhaustive groups. If $\boldsymbol{\beta}_g$ denotes the vector of regression coefficients pertinent to group g, then

the group Lasso

$$\rho \sum_{g \in G} w_g ||\boldsymbol{\beta}_g||_2$$

represents a reasonable penalty that tends to select model parameters in groups rather than individually. Here, the weight w_g is typically chosen as the square root $\sqrt{|g|}$ of the group size $|g|$. If currently $\boldsymbol{\beta}_g = \mathbf{0}$, then the Euclidean penalty $||\boldsymbol{\beta}_g||_2$ reduces to a Lasso penalty as $\beta_j \in g$ is updated. Once one β_j in g lifts off 0, the remaining β_j in g lift off 0 more easily.

Ethnic admixture: percentage of genes from specific founder populations in individuals whose ancestry can be traced to two or more distinct populations

Zhou et al. (2010) apply a combination of Lasso and Euclidean penalties in their regression analysis of breast cancer data. Another instructive example is Wu & Lange's (2008) earlier application of the Lasso to studies of celiac disease. The latter example illustrates an advantage of continuous model selection over traditional GWAS analysis in which models are built up by testing one SNP at a time. The data, originally published by van Heel et al. (2007), consist of 2,200 subjects genotyped for more than 300,000 SNPs. Both Lasso penalized and ordinary univariate logistic regression reveal a strong association between celiac disease and SNPs in the major histocompatibility complex class II region (human chromosome segment 6p21.3). The difference between the two approaches can be seen in testing for two-way gene-by-gene interactions. Using a relatively weak penalty that allows 50 predictors to enter the model, Wu & Lange (2008) find evidence for four gene-by-gene interactions among these predictors. Two of the four interactions involve SNPs whose marginal p-values were not deemed significant at a genomewide threshold of 10^{-7}.

In practice, the Lasso shrinks as well as selects. Severe shrinkage encourages false positives to enter a model to compensate. Statisticians have suggested two remedies. One is to substitute nonconvex penalties for the Lasso. For example, the minimax concave penalty (MCP) (Zhang 2010)

$$\rho p(t) = \rho \int_0^{|t|} \left(1 - \frac{s}{\rho\gamma}\right)_+ \mathrm{d}s$$

starts out at $t = 0$ with slope ρ and gradually transitions to slope 0 at $t = \rho\gamma$. The beauty of this penalty is that it can be majorized—as discussed in Section 3—by a v-shaped function very much like the Lasso's absolute value function. Thus, with minor differences, the coordinate descent algorithm carries over (Breheny & Huang 2011, Mazumder et al. 2011). Model selection is achieved without severe shrinkage, and inference in GWAS improves (Hoffman et al. 2013). The second remedy, stability selection, weeds out false positives by looking for consistent predictor selection across random halves of the data (Alexander & Lange 2011b, Meinshausen & Bühlmann 2010).

2.2. Ethnic Admixture

Population stratification is a potential confounding factor in genetic association studies. Fortunately, estimated ancestries derived from multilocus genotype data can serve as covariates in correcting for population stratification. Because it relies on Bayesian MCMC, the popular program STRUCTURE is intolerably slow (Pritchard et al. 2000). Alternatives such as EIGENSTRAT (part of the EIGENSOFT package) deliver principal components (Price et al. 2006); these flag quality control issues and cryptic relatedness in addition to population stratification. We now discuss a fast, model-based method, which is embodied in the program ADMIXTURE (Alexander & Lange 2011a, Alexander et al. 2009). As its name implies, ADMIXTURE delivers admixture fractions, which are easier to interpret than principal component scores.

Admixture fractions have been widely used to infer aspects of history from genetic data. Some examples are (*a*) the evolutionary histories of populations, such as Jews (Behar et al. 2010), and of individuals, such as an ancient Palaeo-Eskimo (Rasmussen et al. 2010); (*b*) the migratory patterns of hunter-gatherers (Henn et al. 2011); (*c*) the breeding histories of domesticated plants and animals

(Alhaddad et al. 2013, Kijas et al. 2012, Morris et al. 2013); and (*d*) the linguistic stratification in populations (Pagani et al. 2012). Admixture fractions have been used as covariates in several gene mapping analyses, including studies of pulmonary function (Kumar et al. 2010), neuroblastoma (Latorre et al. 2012), and systemic lupus erythematosus (Sánchez et al. 2012).

In the standard population admixture model, population k contributes a fraction w_{ik} of individual i's genome. The reference allele at SNP j has frequency f_{kj} in population k. In unsupervised learning, both the matrices $\mathbf{W} = (w_{ik})$ and $\mathbf{F} = (f_{kj})$ are unknown. In supervised learning, $\mathbf{F}$ is known. The model makes the reasonable assumption that gametes combine randomly and the dubious assumption that all SNPs are inherited independently. Let y_{ij} represent the observed number of copies of the reference allele at marker j of person i. Thus, y_{ij} equals 0, 1, or 2, and the log likelihood of the data is

$$L(\mathbf{W}, \mathbf{F}) = \sum_i \sum_j \left\{ y_{ij} \ln \left[\sum_k w_{ik} f_{kj} \right] + (2 - y_{ij}) \ln \left[\sum_k w_{ik}(1 - f_{kj}) \right] \right\}. \qquad 2.$$

Since all parameters play a role in inference, no penalties are added to the log likelihood.

There are some obvious hindrances to maximizing $L(\mathbf{W}, \mathbf{F})$. Given I unrelated sample people, J SNPs, and K ancestral populations, the parameter matrices $\mathbf{W} = \{w_{ik}\}$ and $\mathbf{F} = \{f_{kj}\}$ have dimensions $I \times K$ and $K \times J$, respectively, for a total of $IK + KJ$ parameters. The modest choices $I = 1{,}000, J = 10{,}000$, and $K = 3$ yield 33,000 parameters to estimate. Thus, the sheer number of parameters makes Newton's method and scoring infeasible. The storage required for the Hessian matrix is prohibitively large, and the required matrix inversion is intractable. Moreover, the log likelihood has at least $K!$ equivalent global maxima and is subject to the bounds $0 \leq f_{kj} \leq 1$ and $w_{ik} \geq 0$ and the equality constraint $\sum_k w_{ik} = 1$.

Block ascent is an effective strategy for maximizing the log likelihood given by Equation 2. Block ascent alternates between updating the $\mathbf{W}$ and $\mathbf{F}$ matrices. The log likelihood is concave both in $\mathbf{W}$ when $\mathbf{F}$ is fixed and in $\mathbf{F}$ when $\mathbf{W}$ is fixed. In the $\mathbf{W}$ updates, the admixture proportions for each individual i are optimized separately. In the $\mathbf{F}$ updates, the allele frequencies for each SNP are optimized separately. The updates of the f_{kj} are exact. The updates of the w_{ik} are found iteratively by sequential quadratic programming; this tactic repeatedly maximizes the second-order Taylor expansion of $L(\mathbf{W}, \mathbf{F})$ around the current parameter vector. Without constraints, sequential quadratic programming coincides with Newton's method. Block ascent can be accelerated by a generic secant method (Zhou et al. 2011a). Standard errors are calculated via the parametric bootstrap. The ADMIXTURE program implementing block ascent is three orders of magnitude faster than STRUCTURE. To our knowledge, no one has carefully compared the suitability of admixture coefficients versus principal components in analyzing GWAS data.

3. MM ALGORITHMS

The MM algorithm is a principle for constructing optimization algorithms (Hunter & Lange 2004, Lange 2010, Wu & Lange 2010). The basic idea is to convert a complex optimization problem into a sequence of simpler ones. In minimization, the MM principle majorizes the objective function $f(\mathbf{x})$ by a surrogate function $g(\mathbf{x}|\mathbf{x}_n)$ anchored at the current point $\mathbf{x}_n$. Majorization combines the tangency condition $g(\mathbf{x}_n|\mathbf{x}_n) = f(\mathbf{x}_n)$ and the domination condition $g(\mathbf{x}|\mathbf{x}_n) \geq f(\mathbf{x})$ for all $\mathbf{x}$. The next iterate of the MM algorithm $\mathbf{x}_{n+1}$ is defined to minimize $g(\mathbf{x}|\mathbf{x}_n)$. Since

$$f(\mathbf{x}_{n+1}) \leq g(\mathbf{x}_{n+1}|\mathbf{x}_n) \leq g(\mathbf{x}_n|\mathbf{x}_n) = f(\mathbf{x}_n),$$

the MM iterates generate a descent algorithm that drives the objective function downhill. Strictly speaking, this descent property depends only on decreasing $g(\mathbf{x}|\mathbf{x}_n)$, not on minimizing it.

Constraint satisfaction is automatically enforced in finding $\mathbf{x}_{n+1}$. Under appropriate regularity conditions, an MM algorithm is guaranteed to converge to a local minimum of the objective function (Lange 2010). In maximization, we first minorize and then maximize. Thus, the acronym MM does double duty in the forms majorize-minimize and minorize-maximize.

When successful, the MM algorithm simplifies optimization by (*a*) separating the variables of a problem, (*b*) avoiding large matrix inversions, (*c*) linearizing a problem, (*d*) restoring symmetry, (*e*) dealing gracefully with equality and inequality constraints, and (*f*) making a nondifferentiable problem smooth. The art in devising an MM algorithm lies in choosing a tractable surrogate function $g(\mathbf{x}|\mathbf{x}_n)$ that hugs the objective function $f(\mathbf{x})$ as tightly as possible.

The majorization relation between functions is closed under the formation of sums, nonnegative products, limits, and composition with an increasing function. These rules allow one to simplify complicated objective functions one piece at a time. Skill in dealing with inequalities is crucial in constructing majorizations. Classical inequalities such as Jensen's inequality, the information inequality, the arithmetic mean–geometric mean inequality, and the Cauchy–Schwartz inequality prove useful in many problems. The supporting hyperplane property of a convex function and the quadratic upper bound principle of Böhning & Lindsay (1988) also have many applications.

All expectation-maximization (EM) algorithms are also MM algorithms (Lange 2010, McLachlan & Krishnan 2007). Minorization is achieved via the information inequality by contrasting observed data with complete data. Since the basic inequality is given, the difficulty in constructing EM algorithms lies in identifying the complete data and calculating required conditional expectations. Statistical geneticists derived gene- and haplotype-counting algorithms long before the EM principle fully justified their use as ascent algorithms (Lange 2002, Smith 1957).

3.1. Matrix Completion and Genotype Imputation

SNP imputation can be viewed as a matrix completion problem. In the machine learning community, matrix completion is a popular and effective imputation tool in many application domains outside of genetics (Cai et al. 2010, Candès & Tao 2010, Chen et al. 2012a, Mazumder et al. 2010). This tool can recover an entire matrix when only a small portion of its entries is actually observed. In the pursuit of parsimony, matrix completion seeks the simplest matrix that is consistent with the observed entries. This criterion conveniently translates into searching for a low-rank matrix with a small squared error difference over the observed entries. The celebrated Netflix Challenge represented a typical application to recommender systems (ACM-SIGKDD & Netflix 2007, ACM-SIGKDD 2007). The goal of the Netflix Challenge was to impute a $480{,}189 \times 17{,}770$ matrix in which nearly 99% of the original entries were missing.

Imputing missing genotypes shares many features with the Netflix Challenge. Genotypes can be coded as 0, 1, or 2 by counting reference alleles, such as the least frequent allele, and entering the counts into a matrix with rows labeled by individuals and columns labeled by SNPs. Despite its purely empirical nature, matrix completion can achieve good accuracy with relatively little computational effort. In SNP imputation, matrix completion is performed via a window that is slid along the genome (Chi et al. 2013). Its success is predicated on a low-rank structure in the data. LD ensures the validity of this assumption over short windows of contiguous SNPs.

Let $\mathbf{Y}$ denote a partially observed matrix and Δ denote the set of index pairs (i, j) with y_{ij} observed. Matrix completion minimizes the criterion

$$f(\mathbf{X}) = \frac{1}{2} \sum_{(i,j)\in\Delta} (y_{ij} - x_{ij})^2 + \rho \sum_k \sigma_k \qquad 3.$$

for a compatible matrix $\mathbf{X} = (x_{ij})$ with singular values σ_k. The positive tuning constant ρ determines the strength of the rank penalty. The singular values appear in the singular value decomposition (SVD)

$$\mathbf{X} = \sum_i \sigma_i \mathbf{u}_i \mathbf{v}_i^t.$$

This sum-of-outer-products form of the SVD invokes an orthonormal collection of left singular vectors $\mathbf{u}_i$, a corresponding orthonormal collection of right singular vectors $\mathbf{v}_i$, and a descending sequence of nonnegative singular values σ_i. Equivalently, one can write the SVD of $\mathbf{X}$ in factored form as $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^t$ for orthogonal matrices $\mathbf{U}$ and $\mathbf{V}$ and a rectangular diagonal matrix $\mathbf{\Sigma}$.

The nuclear norm $||\mathbf{X}||_{\text{nuc}} = \sum_k \sigma_k$ plays the same role in low-rank matrix approximation that the ℓ_1 norm $||\mathbf{b}||_1 = \sum_k |b_k|$ plays in sparse regression. To represent the matrix completion criterion from Equation 3 more succinctly, one can introduce the Frobenius norm

$$||\mathbf{M}||_F = \sqrt{\text{tr}(\mathbf{M}\mathbf{M}^t)} = \sqrt{\sum_i \sum_j m_{ij}^2}$$

induced by the trace inner product $\text{tr}(\mathbf{M}\mathbf{N}^t)$ and the projection operator $P_\Delta(\mathbf{Y})$ with entries

$$P_\Delta(\mathbf{Y}) = \begin{cases} y_{ij} & (i,j) \in \Delta \\ 0 & (i,j) \notin \Delta. \end{cases}$$

Using this notation, the matrix completion criterion in Equation 3 becomes

$$f(\mathbf{X}) = \frac{1}{2}||P_\Delta(\mathbf{Y}) - P_\Delta(\mathbf{X})||_F^2 + \rho||\mathbf{X}||_{\text{nuc}}.$$

After estimating the optimal $\mathbf{X}$, genotypes can be imputed by clustering the entries of each column of $\mathbf{X}$ into three groups corresponding to the coded genotypes 0, 1, and 2.

The MM algorithm allows for the restoration of the symmetry lost in the missing entries (Mazumder et al. 2010). Suppose $\mathbf{X}_n$ is our current approximation of $\mathbf{X}$. We simply replace a missing entry y_{ij} of $\mathbf{Y}$ where $(i,j) \notin \Delta$ with the corresponding entry x_{nij} of $\mathbf{X}_n$ and add the term $\frac{1}{2}(x_{nij} - x_{ij})^2$ to the criterion in Equation 3. Since the added terms majorize 0, they create a legitimate surrogate function and lead to an MM algorithm. The problem can be rephrased in matrix terms by defining the orthogonal complement operator $P_\Delta^\perp(\mathbf{Y})$ via $P_\Delta^\perp(\mathbf{Y}) + P_\Delta(\mathbf{Y}) = \mathbf{Y}$. Then, the matrix $\mathbf{Z}_n = P_\Delta(\mathbf{Y}) + P_\Delta^\perp(\mathbf{X}_n)$ temporarily completes $\mathbf{Y}$ and yields the surrogate function

$$\begin{aligned} g(\mathbf{X}|\mathbf{X}_n) &= \frac{1}{2}||\mathbf{Z}_n - \mathbf{X}||_F^2 + \rho||\mathbf{X}||_{\text{nuc}} \\ &= \frac{1}{2}||\mathbf{Z}_n||_F^2 - \text{tr}(\mathbf{Z}_n\mathbf{X}^t) + \frac{1}{2}||\mathbf{X}||_F^2 + \rho||\mathbf{X}||_{\text{nuc}}. \end{aligned}$$

To make further progress, recall that the Frobenius norm is invariant under left and right multiplication of its argument by an orthogonal matrix. Thus, $||\mathbf{X}||_F^2 = \sum_k \sigma_k^2$ depends only on the singular values of $\mathbf{X}$. Majorizing the inner product $-\text{tr}(\mathbf{Z}_n\mathbf{X}^t)$ is more subtle. Fortunately, one can apply a matrix analog of the Cauchy–Schwarz inequality. Fan's inequality (Borwein & Lewis 2006) says that

$$\text{tr}(\mathbf{Z}_n\mathbf{X}^t) \leq \sum_k \omega_k \sigma_k$$

for the ordered singular values ω_k of $\mathbf{Z}_n$. Equality is attained in Fan's inequality if and only if the right and left singular vectors for the two matrices $\mathbf{Z}_n$ and $\mathbf{X}$ coincide. Thus, in minimizing $g(\mathbf{X}|\mathbf{X}_n)$, we can assume that the singular vectors of $\mathbf{X}$ coincide with those of $\mathbf{Z}_n$ and rewrite the

surrogate function as

$$g(\mathbf{X}|\mathbf{X}_n) = \frac{1}{2}\sum_k \omega_k^2 - \sum_k \omega_k\sigma_k + \frac{1}{2}\sum_k \sigma_k^2 + \rho\sum_k \sigma_k$$
$$= \frac{1}{2}\sum_k (\omega_k - \sigma_k)^2 + \rho\sum_k \sigma_k.$$

Standard calculus arguments demonstrate that the shrunken singular values

$$\sigma_k = \max\{\omega_k - \rho, 0\}$$

are optimal. In practice, the full SVD of $\mathbf{Z}_n$ need not be extracted. Rather, only the singular values $\omega_k > \rho$ are actually relevant in constructing $\mathbf{X}_{n+1}$.

Hardy–Weinberg equilibrium: an equilibrium state characterized by stable allele and genotype population frequencies; also a rule to calculate genotype frequencies from allele frequencies assuming a random union of gametes

The above example teaches several timely lessons (Chi et al. 2013). First, although matrix completion largely ignores the underlying genetics, it imputes genotypes almost as accurately as the best hidden Markov models. Second, matrix completion is more than an order of magnitude faster than competing imputation techniques because it focuses on relatively short windows and exploits well-established algorithms for extracting singular values and vectors. Third, the versatility and simplicity of the MM principle are on vivid display. Fourth, more exotic convex programming methods that rely on Nesterov acceleration (Beck & Teboulle 2009, Nesterov 2007) can achieve even more impressive speedups. Fifth, although the usual assumption of data missing completely at random fails, matrix completion delivers good results. (In practice, most of the missing entries of the genotype matrix occur because study subjects are genotyped on different platforms that have different SNP sets.) Sixth, awareness of developments in data mining, even outside one's narrow application area, pays rich dividends.

3.2. Matrix Completion with Sequence Data

Matrix completion can also be applied to low-coverage sequence data, provided that read counts are converted into expected dosages. A simple binomial model enables this conversion (Pasaniuc et al. 2012, Sampson et al. 2011). Let G_{ij} denote the latent genotype at SNP i for individual j, A and B denote the major and minor alleles at SNP i, and $R_{ij} = (a_{ij}, b_{ij})$ denote the read count pair for j over A and B at SNP i, respectively. For a uniform read error of ε and a fixed value of the coverage $n_{ij} = a_{ij} + b_{ij}$, one has

$$\Pr[R_{ij} = (a_{ij}, b_{ij})|G_{ij} = A/A] = \binom{n_{ij}}{a_{ij}}(1-\varepsilon)^{a_{ij}}\varepsilon^{b_{ij}},$$

$$\Pr[R_{ij} = (a_{ij}, b_{ij})|G_{ij} = A/B] = \binom{n_{ij}}{a_{ij}}(1/2)^{n_{ij}},$$

$$\Pr[R_{ij} = (a_{ij}, b_{ij})|G_{ij} = B/B] = \binom{n_{ij}}{a_{ij}}\varepsilon^{a_{ij}}(1-\varepsilon)^{b_{ij}}.$$

To convert these read counts into posterior expectations, one can reasonably impose Hardy–Weinberg priors

$$\Pr(G_{ij} = A/A) = p_A^2,$$
$$\Pr(G_{ij} = A/B) = 2p_Ap_B,$$
$$\Pr(G_{ij} = B/B) = p_B^2,$$

where p_A and p_B are the estimated allele frequencies from a study or reference panel such as the 1000 Genomes Project (1000 Genomes Proj. Consort. 2010). This leads to the posterior probabilities

$$q_{A/A} = \Pr[G_{ij} = A/A | R_{ij} = (a_{ij}, b_{ij})] = \frac{(1-\varepsilon)^{a_{ij}} \varepsilon^{b_{ij}} p_A^2}{Z},$$
$$q_{A/B} = \Pr[G_{ij} = A/B | R_{ij} = (a_{ij}, b_{ij})] = \frac{2(1/2)^{n_{ij}} p_A p_B}{Z},$$
$$q_{B/B} = \Pr[G_{ij} = B/B | R_{ij} = (a_{ij}, b_{ij})] = \frac{\varepsilon^{a_{ij}} (1-\varepsilon)^{b_{ij}} p_B^2}{Z},$$

with normalizing constant

$$Z = (1-\varepsilon)^{a_{ij}} \varepsilon^{b_{ij}} p_A^2 + 2(1/2)^{n_{ij}} p_A p_B + \varepsilon^{a_{ij}} (1-\varepsilon)^{b_{ij}} p_B^2.$$

Finally, if A is the reference allele, then the posterior mean dosage can be expressed as the sum $x_{ij} = 2q_{A/A} + q_{A/B}$ and fed into the matrix completion algorithm.

Copy number variation (CNV): tendency of genomic regions to be duplicated or lost, leading to variation among individuals

3.3. The Fused Lasso and Copy Number Variation

Copy number variants (CNVs) exist throughout the human genome and range in size from a few kilobases to a few megabases. For a given SNP with alleles A and B, genotyping platforms typically record the total DNA signal on a log scale, as well as the fraction of the signal attributed to the B allele. Such data allow one to impute copy number along the genome. Normal DNA generates a copy number of 2, deletions generate copy numbers of 0 or 1, and insertions generate copy numbers of 3 or more. Here we consider a simplified version of the problem that employs only signal intensity. Let y_i and β_i denote the observed and theoretical signal intensities at SNP i. With m SNPs, the fused Lasso model (Tibshirani et al. 2005) for estimating the parameter vector $\boldsymbol{\beta}$ minimizes the criterion

$$f(\boldsymbol{\beta}) = \frac{1}{2}\sum_{i=1}^{m}(y_i - \beta_i)^2 + \rho_1 \sum_{i=1}^{m}|\beta_i| + \rho_2 \sum_{i=2}^{m}|\beta_i - \beta_{i-1}|, \qquad 4.$$

where the first penalty pulls $\hat{\beta}_i$ toward 0, the standardized value of y_i for a copy number of 2, and the second penalty controls the number of jumps between successive piecewise constant segments. The tuning constants ρ_1 and ρ_2 determine the strength of these penalties.

Unfortunately, this twist on the standard Lasso penalty stymies coordinate descent. To construct an MM algorithm, we replace the nondifferentiable function $|x|$ by $||x||_{2,\varepsilon} = \sqrt{x^2 + \varepsilon}$ for small $\varepsilon > 0$. The concavity of the function $u \mapsto \sqrt{u + \varepsilon}$ on the interval $[0, \infty)$ yields the majorization

$$||x||_{2,\varepsilon} \leq ||x_n||_{2,\varepsilon} + \frac{1}{2||x_n||_{2,\varepsilon}}[x^2 - x_n^2].$$

Substituting $||x||_{2,\varepsilon}$ for $|x|$ in the criterion given by Equation 4 leads to the surrogate function

$$g(\boldsymbol{\beta}|\boldsymbol{\beta}_n) = \frac{1}{2}\sum_{i=1}^{m}(y_i - \beta_i)^2 + \frac{\rho_1}{2}\sum_{i=1}^{m}\frac{\beta_i^2}{||\beta_{ni}||_{2,\varepsilon}} + \frac{\rho_2}{2}\sum_{i=2}^{n}\frac{(\beta_i - \beta_{i-1})^2}{||\beta_{ni} - \beta_{n,i-1}||_{2,\varepsilon}} + c_n$$

depending on an irrelevant constant c_n. The surrogate can be written as the quadratic function

$$g(\boldsymbol{\beta}|\boldsymbol{\beta}_n) = \frac{1}{2}\boldsymbol{\beta}^t \mathbf{A}_n \boldsymbol{\beta} - \mathbf{b}_n^t \boldsymbol{\beta} + c_n,$$

where $\mathbf{A}_n$ is a tridiagonal positive definite matrix. The minimum occurs at the point $\boldsymbol{\beta} = \mathbf{A}_n^{-1}\mathbf{b}_n$. The Thomas algorithm (also known as the tridiagonal matrix algorithm) solves the linear system

$\mathbf{A}_n\boldsymbol{\beta} = \mathbf{b}_n$ in just $O(m)$ operations. Overall, the MM algorithm with this surrogate converges rapidly for tens of thousands of SNPs (Zhang et al. 2010).

This application of the MM algorithm develops three key ideas (Zhang et al. 2012). The first is to pose the estimation problem in terms of penalized least squares. The second is to approximate the absolute value function by a smooth function that can be majorized by a quadratic. The third is to recognize the importance of the Thomas algorithm. In practice, the tuning constants ρ_1 and ρ_2 are chosen by a judicious combination of statistical theory and cross validation. Readers are referred to Zhang et al. (2012) for elaborations of this model and alternative treatments.

3.4. Haplotyping

Haplotyping goes beyond genotype matrix completion and delivers the maternal and paternal phase of each observed or imputed genotype. Two key ideas dominate the literature on haplotyping (Ayers & Lange 2008, Browning & Browning 2007, Chen et al. 2012b, Howie et al. 2012, Li et al. 2010, Scheet & Stephens 2006, Stephens et al. 2001, Williams et al. 2012). Foremost is the notion of LD. Nature and population history strictly limit the number of possible haplotypes in a short genomic region. The resulting LD represents a failure of the product rule for independent events in computing haplotype frequencies from allele frequencies. Consequently, imposing parsimony is absolutely crucial in haplotyping. Most current haplotyping methods achieve parsimony indirectly through hidden Markov models. Here we briefly explore direct penalization. The second key idea, the application of guide haplotypes, is also critically important. The accumulated data from the HapMap Project (Int. HapMap Consort. 2003) and the 1000 Genomes Consortium (1000 Genomes Proj. Consort. 2010) contain haplotypes for literally thousands of ethnically diverse people. Thus, in each short genomic region, the universe of possible haplotypes is well known.

The traditional EM algorithm for haplotype frequency estimation (Excoffier & Slatkin 1995, Long et al. 1995) is an MM algorithm that operates over a narrow genomic window. If $\mathbf{q}$ is the vector of haplotype frequencies and H_i is the set of ordered haplotype pairs (k, l) consistent with subject i's observed multilocus genotype, then i's likelihood can be written as

$$\ell_i(\mathbf{q}) = \sum_{(k,l)\in H_i} q_k q_l.$$

The full log likelihood across all independent samples equals

$$L(\mathbf{q}) = \sum_i \ln \ell_i(\mathbf{q}).$$

One can encourage parsimony by subtracting from $L(\mathbf{q})$ a penalty that tends to eliminate haplotypes with low explanatory power (Ayers & Lange 2008). The penalty is defined by a threshold δ, a tuning constant ρ that scales the strength of the penalty, and the penalty function

$$p(q) = \begin{cases} q & q \le \delta \\ \delta & q > \delta. \end{cases}$$

In the penalized MM algorithm, one estimates the parameter vector $\mathbf{q}$ by maximizing the objective function

$$f(\mathbf{q}) = L(\mathbf{q}) - \rho \sum_j p(q_j). \qquad 5.$$

The concavity of the logarithm function leads to the minorization

$$L(\mathbf{q}) \geq \sum_i \sum_{(k,l)\in H_i} \frac{q_{nk}q_{nl}}{\ell_i(\mathbf{q}_n)} \ln\left[\frac{\ell_i(\mathbf{q}_n)}{q_{nk}q_{nl}} q_k q_l\right]$$
$$= \sum_k c_{nk} \ln q_k + c_{n0}$$

invoking the constants

$$c_{nk} = \sum_i \sum_l \left[1_{(k,l)\in H_i} + 1_{(l,k)\in H_i}\right] \frac{q_{nk}q_{nl}}{\ell_i(\mathbf{q}_n)},$$
$$c_{n0} = \sum_i \sum_{(k,l)\in H_i} \frac{q_{nk}q_{nl}}{\ell_i(\mathbf{q}_n)} \ln\left[\frac{\ell_i(\mathbf{q}_n)}{q_{nk}q_{nl}}\right].$$

The penalty $p(q_j)$ is majorized by the linear function q_j when $q_{nj} \leq \delta$ and by the constant δ when $q_{nj} > \delta$. Multiplying the penalty majorization by $-\rho$ gives a minorization $-\rho p(q_j)$. Overall, we derive the minorization

$$f(\mathbf{q}) \geq \sum_j c_{nj} \ln q_j + c_{n0} - \rho \left(\sum_{j:q_{nj}<\delta} q_j + \sum_{j:q_{nj}\geq\delta} \delta \right)$$

of the objective function given by Equation 5. The maximization step of the MM algorithm involves solving a sequence of quadratic equations that respect the constraints $q_j \geq 0$ and $\sum_j q_j = 1$. Details appear in Ayers & Lange (2008). Notably, the explicit MM updates are only slightly more complicated than the standard EM updates.

Haplotyping relies on Bayes' rule and the estimated frequencies of the reference haplotypes. Suppose (k_m, l_m) is the ordered genotype at SNP m derived from the haplotype pair (k, l). The posterior probability of i having the ordered genotype (s, t) at m is just the ratio

$$\frac{1}{\ell_i(\mathbf{q})} \sum_{(k,l):(k_m,l_m)=(s,t)} q_k q_l.$$

If hard imputations are desired, then the unordered genotype with the maximum posterior probability is pertinent. In practice, it is convenient to impute haplotypes in the middle third of the current window. Advancing to the next window requires deleting the left third and adding a new right third so that the former right third occupies the middle of the current window. The process is highly parallel and can be accomplished by the simple numerical processors of the graphics processing unit (GPU) in most modern desktop and laptop computers (Chen et al. 2012b). Computations that once took months can now be done in hours on a typical computer. The value of the tuning constant ρ is adapted to the local level of LD by applying cross validation to randomly deleted entries in the two extreme thirds of a window. This calibration step makes penalized haplotyping comparable in accuracy to hidden Markov haplotyping.

4. GENE MAPPING VIA PEDIGREES

4.1. Estimation of Relatedness

Traditionally, degrees of relatedness between pairs of individuals have been deduced from pedigree graphs (Jacquard 1970, Lange 2002). As stated in Section 1, it is desirable to estimate the various coefficients of relatedness empirically from dense genotyping data. **Figure 1** depicts the nine condensed identity states established by Jacquard (1970). In the figure, two genes (dots) at the same locus are connected by a line if they are identical by descent (IBD), meaning both are physical

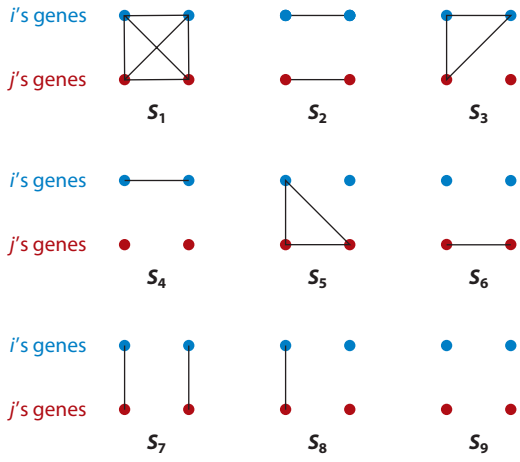

Figure 1

The nine condensed identity states presented in Jacquard (1970) illustrate the possible relationships between two individuals, *i* and *j*. For each individual, the two genes at an arbitrary locus are depicted by dots: blue for *i* and red for *j*. Lines connecting two dots indicate genes that are identical by descent; that is, both genes are inherited copies of the same ancestral gene. In the absence of inbreeding, only the last three states (S_7, S_8, and S_9) are possible.

copies of the same ancestral gene. Relatedness is quantified by the probabilities Δ_1 through Δ_9 of the nine states at a randomly chosen locus. The first six states are impossible unless one or both individuals are inbred. Although graph-traversing algorithms for computing the Δ_i values exist, these algorithms are complicated to program and rely on the fidelity of recorded pedigrees (Lange 2002).

The kinship coefficient

$$\mathbf{\Phi}_{ij} = \Delta_1 + \frac{1}{2}(\Delta_3 + \Delta_5 + \Delta_7) + \frac{1}{4}\Delta_8$$

between two individuals *i* and *j* is the single most valuable coefficient of relatedness. It can be interpreted as the probability that a randomly sampled gene from individual *i* is IBD to a randomly sampled gene from individual *j* at the same locus. When $i = j$, the sampling is done with replacement. Thus, in the absence of inbreeding, $\mathbf{\Phi}_{ii} = \frac{1}{2}$, and $\mathbf{\Phi}_{ij} = \frac{1}{4}$ for a sibling or parent-offspring pair.

Condensed identity coefficients can be estimated in several ways (Boehnke & Cox 1997; Day-Williams et al. 2011; Thompson 1974, 1975). One of the simplest ways of estimating kinship coefficients minimizes the expected number of identical by state (IBS) matches between individuals *i* and *j* under random gene sampling. If *m* SNPs are sampled and the frequency of the reference allele for SNP *k* is p_k, then the expected number of IBS matches equals

$$e_{ij} = \sum_{k=1}^{m} \{\mathbf{\Phi}_{ij} + (1 - \mathbf{\Phi}_{ij})[p_k^2 + (1 - p_k)^2]\}.$$

The first term in the summation accounts for matches that are IBD at SNP *k*, whereas the second term accounts for matches that are IBS but not IBD. Solving for $\mathbf{\Phi}_{ij}$ gives

$$\mathbf{\Phi}_{ij} = \frac{e_{ij} - \sum_{k=1}^{m}[p_k^2 + (1 - p_k)^2]}{m - \sum_{k=1}^{m}[p_k^2 + (1 - p_k)^2]}. \qquad 6.$$

In practice, the method of moments formula given by Equation 6 is implemented by equating e_{ij} to the expected number of IBS matches over all m SNPs conditional on observed genotypes (see Day-Williams et al. 2011 for details). Yang et al. (2010) independently pursued a similar problem.

Dense SNP genotyping raises the possibility of estimating identity coefficients from inferred haplotypes. If one accepts the premise that haplotype identity is equivalent to being IBD, then locally along the genome one can immediately deduce the condensed identity state of a pair of individuals i and j. Suppose in a narrow window surrounding a designated SNP there are h possible haplotypes labeled $1, \ldots, h$. For example, if $h = 10$, i has haplotype pair (7, 2), and j has haplotype pair (7, 7), then inspection of **Figure 1** shows that i and j can be assigned to identity state S_5. Globally, Δ_i can reasonably be set to the fraction of the surveyed SNPs that are locally in state S_i.

Polygenes: a large collection of genes that each have a small, additive effect on a trait value but that together account for a nontrivial amount of the trait variation among individuals

4.2. Variance Components Models

Association testing is much simpler with case-control or random sample data than with pedigree data. From its inception, linkage analysis has been forced to confront the complications of pedigrees (Elston & Stewart 1971). Fortunately, Gaussian pedigree models combine some of the best features of linkage analysis and association testing for quantitative traits. These mixed-effects models properly account for polygenic background and correlated environments, and they encourage the analysis of multivariate traits. If score tests are substituted for likelihood ratio tests, then hundreds of thousands of SNPs can be processed in a short amount of time (Lange et al. 2013). Association testing focuses on mean effects; linkage analysis focuses on random effects. Despite the complications of pedigree data, there are two compelling reasons for analyzing pedigrees. First, many pedigrees were collected in the linkage era of gene mapping. For instance, twin registries are full of simple pedigree data. Second, pedigrees tend to concentrate rare variants with major phenotypic effects. The same is true for population isolates in which almost all individuals are related.

The Gaussian pedigree model invokes a multivariate Gaussian distribution to model the vector of observed trait values $\mathbf{y}$ from a pedigree. Under the standard model (Lange 2002), the log likelihood of a pedigree reduces to

$$L = -\frac{1}{2}\ln\det\boldsymbol{\Omega} - \frac{1}{2}(\mathbf{y}-\boldsymbol{\nu})^t\boldsymbol{\Omega}^{-1}(\mathbf{y}-\boldsymbol{\nu}),$$

where $\boldsymbol{\nu}$ is the vector of trait means, and $\boldsymbol{\Omega}$ is the matrix of trait covariances. The trait in question can be univariate or multivariate. Pedigrees are considered independent observational units. For a univariate trait, the covariance matrix is typically parameterized as

$$\boldsymbol{\Omega} = 2\sigma_a^2\boldsymbol{\Phi} + \sigma_d^2\boldsymbol{\Delta}_7 + \sigma_e^2\mathbf{I}, \qquad 7.$$

where $\boldsymbol{\Phi}$ is the global kinship coefficient matrix capturing additive polygenic effects, and $\boldsymbol{\Delta}_7$ is a condensed identity coefficient matrix capturing genetic dominance effects. Individual environmental contributions and trait measurement errors are incorporated via the identity matrix $\mathbf{I}$. Other random effects, such as household effects, can be added as needed. When all individuals are definitely unrelated, the covariance matrix reduces to $\boldsymbol{\Omega} = \sigma_e^2\mathbf{I}$. In addition, if a pedigree structure is unknown or dubious, $\boldsymbol{\Phi}$ can be accurately estimated from dense markers via the method of moments discussed in the previous section. Pedigree clusters can then be constructed from $\boldsymbol{\Phi}$ as connected components of a relationship graph (Day-Williams et al. 2011).

A linear model assumes that $\boldsymbol{\nu} = \mathbf{A}\boldsymbol{\beta}$ for a design matrix $\mathbf{A}$ and vector $\boldsymbol{\beta}$ of regression coefficients. In SNP association testing, one of the predictors is the dosage level of a SNP, namely the imputed count of the number of reference alleles. This count may be fractional. A likelihood ratio or score test is conducted to decide whether the coefficient corresponding to that predictor

is significantly different from 0. The test statistic asymptotically follows a χ^2_1 distribution. Alternatively, one can assign a regression coefficient to each haplotype in a short genomic region and use the haplotype counts as predictors. In this strategy, haplotypes serve as surrogates for untyped causative SNPs in this region. A possible drawback is that an excess of haplotypes obscures inference and leads to tests with too many degrees of freedom.

Quantitative trait locus (QTL): a gene or SNP that codes for or regulates a quantitative trait as opposed to a disease or other categorical trait

Quantitative trait locus (QTL) mapping exploits random effects and avoids the degrees of freedom dilemma (Almasy & Blangero 1998, Haseman & Elston 1972, Hopper & Mathews 1982, Williams & Blangero 1999). One can modify the variance matrix in Equation 7 by adding another term, $\sigma^2_{\text{loc}}\mathbf{\Phi}_{\text{loc}}$, which captures local IBD sharing at the current point along the genome. In practice, $\mathbf{\Phi}_{\text{loc}}$ can be computed as the sum $\frac{1}{4}\sum_j \mathbf{h}_j\mathbf{h}_j^t$ of the outer products of imputed haplotype vectors. The entry h_{jk} of $\mathbf{h}_j$ is 0, 1, or 2, depending on the number of copies of haplotype j carried by individual k. The fact that $\mathbf{\Phi}_{\text{loc}}$ may no longer be block diagonal in the pedigrees of a study complicates both likelihood ratio testing and score testing. However, this numerical objection is balanced by the added power to exploit cryptic relatedness in mapping. This method of QTL mapping via a variance components model also has the advantage that the single parameter alternative hypothesis $\sigma^2_{\text{loc}} > 0$ parsimoniously flags the presence of linkage or association. Overall, tying imputation of $\mathbf{\Phi}_{\text{loc}}$ to haplotyping is a promising strategy, even if software to do so is not yet available. Other extensions to QTL mapping have been recently developed to account for population structure, inbred strains, and additional computational efficiencies (Aulchenko et al. 2007; Blangero et al. 2013; Broman & Sen 2009; Kang et al. 2010; Laird & Lange 2011; Lee et al. 2011; Yang et al. 2010, 2011; Zhou et al. 2012).

5. ASSOCIATION AND RARE VARIANTS

Finally, let us discuss association testing with rare variants. To improve the power to detect rare variant effects in a gene, genomic region, or pathway, statisticians have proposed a number of lumping strategies (Asimit & Zeggini 2010, Asimit et al. 2012b, Bacanu et al. 2012, Bansal et al. 2010, Kiezun et al. 2012). Despite the advantages of aggregation, it is important to realize that a large number of genomic regions will still be tested, and the effect sizes for complex traits are apt to be small. Hence, the multiple testing problem of GWAS remains in play, requiring large sample sizes to reach genomewide significance (Kiezun et al. 2012).

The first rare variant tests simply collapsed all variants within a region (Li & Leal 2008, Morgenthaler & Thilly 2007); later extensions allowed variants to be weighted in accordance with their prior probabilities of being deleterious (Madsen & Browning 2009, Price et al. 2010). These approaches have the drawback of assuming that all rare variants within a region act in the same direction on trait values or disease risks. Consequently, naive collapsing methods lose power in the presence of protective variants (Asimit et al. 2012b). In contrast, overdispersion tests properly account for both protective and deleterious variants (Ionita-Laza et al. 2011, Neale et al. 2011).

The C-alpha test (Neale et al. 2011) compares the observed variance in the dispersion of rare variants between cases and controls to the expected variance under the null hypothesis of no association. Suppose there are m_1 case chromosomes and m_2 control chromosomes with n_{k1} case variants and n_{k2} control variants at site k. Under the null hypothesis of no association, the number of rare variants observed within the cases follows a hypergeometric distribution with success probability $p = m_1/(m_1 + m_2)$ and total trials $n_k = n_{k1} + n_{k2}$. The test statistic is $Z = T/\sqrt{\text{var}(T)}$, where

$$T = \sum_k [(n_{k1} - n_k p)^2 - n_k p(1-p)].$$

Significance is assessed either by permuting the case and control labels or by noting that Z is asymptotically normal with mean 0 and variance 1 given linkage equilibrium. This test is one-sided because dispersion increases under the alternative hypothesis of association.

Epistasis: interactive effects of variants at different loci

We can also construct a likelihood ratio test that is comparable to the C-alpha test. If we assume that the number of variants is binomially distributed, then at locus k, the likelihoods of p_{k1} and p_{k2}, the frequencies of the case and control variants, respectively, are

$$L_{ki}(p_{ki}) = \binom{m_i}{n_{ki}} p_{ki}^{n_{ki}}(1-p_{ki})^{m_i-n_{ki}}$$

for $i = 1$ and 2. The maximum likelihood estimate of p_{ki} is $\hat{p}_{ki} = n_{ki}/m_i$ in each instance. The likelihood $L_k(p_k)$ and estimate $\hat{p}_k$ for the combined sample are similar. The obvious test statistic

$$\sum_k w_k \left[\ln L_{k1}(\hat{p}_{k1}) + \ln L_{k2}(\hat{p}_{k2}) - \ln L_k(\hat{p}_k)\right]$$

invokes weights such as $w_k = 1/\sqrt{4p_k(1-p_k)}$ that emphasize rare variants. Given the small counts of some variants, it would be prudent to assess significance by permutation of case-control labels. One can construct a similar weighted score test that conditions on the number of variants $n_{k1}+n_{k2}$ at each site k by substituting the hypergeometric distribution for the binomial distribution.

The most general rare variant tests rely on mixed-effects regression models (Liu & Leal 2012, Wu et al. 2011). This framework includes linear regression, logistic regression, and indeed any generalized linear model. Mixed-effects models make it easy to adjust for the effects of confounders such as population stratification. The SKAT software (Wu et al. 2011) implements a computationally efficient score test. If y_i denotes the trait value of subject i, then the predicted value of y_i is $\alpha_0+\mathbf{x}_i^t\boldsymbol{\alpha}+\mathbf{G}_i^t\boldsymbol{\beta}$, where x_i collects the nongenetic predictors and $\mathbf{G}_i$ collects the minor allele counts in the genomic region under consideration. The components of $\boldsymbol{\beta}$ are treated as random effects rather than as fixed effects. SKAT further assumes that $\beta_k \sim N(0, w_k\tau)$, where w_k is a locus specific weight and τ is a positive scalar. Finally, the weight w_k is sampled as $\sqrt{w_k} \sim \text{beta}(p_k, a_1, a_2)$, where p_k is the minor allele frequency at locus k, $a_1 \leq 1$, and $a_2 \geq 1$. These assumptions weight rare variants more heavily than common variants (Wu et al. 2011). When there are no covariates and the weights are all set to 1, SKAT and the C-alpha test are equivalent. Significance is assessed using a score test with a kernel that measures the genetic similarity among individuals. By modifying the kernel to allow for covariances among loci, SKAT can be used to test for epistasis.

6. CONCLUSION

Genetics is a rich vein for statistical application and inspiration. The high-dimensional problems encountered in genetics have spurred critical thinking about data mining (Dziuda 2010, Shah & Kusiak 2004), false discovery rates (Efron 2010, Storey & Tibshirani 2003), network analysis (Horvath 2011), and many other subjects. The trendy but nebulous term "big data" conveys some of the fear of being buried under the crush of genomics data. We have outlined a few novel computational tools that undergird genomic data analysis. As whole-genome sequence, expression, and methylation data become more widely available, statisticians will have ample opportunities to apply these tools and to devise new ones. The scale and complexity of genetic studies have reached the point at which data storage and retrieval are impeding progress. Authorship lists for statistical genetics articles rival in length those for particle physics articles. Nevertheless, opportunities abound for creativity in statistical inference and algorithm construction. Finding the right balance between model accuracy and computational feasibility has been and will continue to be the primary

challenge facing statisticians. A false dichotomy between theory and application should be avoided here, as elsewhere, at all costs.

Let us venture a few guesses about the future of the intertwined fields of statistics and genetics. The trends in modeling, penalization, and optimization featured in this review will continue to unfold. More ambitious network models that reflect both dynamics and connectivity will appear. The quality, extent, and number of genomic databases will increase. Thorough catalogs of the mutations that occur in Mendelian disease genes will be built, and the extent of gene regulation will be revealed. Statistics will see a steady progression toward nonconvex penalties, online algorithms, and a more productive merger of the frequentist and Bayesian paradigms. Unless quantum computing becomes practical, parallel processing remains our best hope for handling big data. Algorithm development must keep in mind hardware limitations and opportunities. The pressure to translate genetic discoveries into pharmaceuticals and personalized medicine will persist. All of these trends create enormous opportunities for statisticians. Communities that nurture statistics, computing, and basic biological research will thrive. Although many basic scientific principles underlying genetics have been discovered, translating these principles into action will easily occupy science, medicine, and humankind for the next century.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

Our original work described in this review was supported by NIH grants from the National Human Genome Research Institute (HG006139) and the National Institute of General Medical Sciences (GM103774 and GM053275).

LITERATURE CITED

1000 Genomes Proj. Consort. 2010. A map of human genome variation from population-scale sequencing. *Nature* 467:1061–73

ACM-SIGKDD. 2007. *KDD Cup overview: consumer recommendations*. **http://www.kdd.org/kdd-cup-2007-consumer-recommendations**

ACM-SIGKDD, Netflix. 2007. *Proceedings of the KDD Cup and Workshop 2007*. New York: ACM. **http://www.cs.uic.edu/~liub/KDD-cup-2007/proceedings.html**

Alexander DH, Lange K. 2011a. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinforma.* 12:246

Alexander DH, Lange K. 2011b. Stability selection for genome-wide association. *Genet. Epidemiol.* 35:722–28

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 19:1655–64

Alhaddad H, Khan R, Grahn RA, Gandolfi B, Mullikin JC, et al. 2013. Extent of linkage disequilibrium in the domestic cat, *Felis silvestris catus*, and its breeds. *PLoS ONE* 8:e53537

Almasy L, Blangero J. 1998. Multipoint quantitative-trait linkage analysis in general pedigrees. *Am. J. Hum. Genet.* 62:1198–211

Asimit J, Day-Williams A, Zgaga L, Rudan I, Boraska V, Zeggini E. 2012a. An evaluation of different meta-analysis approaches in the presence of allelic heterogeneity. *Eur. J. Hum. Genet.* 20:709–12

Asimit J, Zeggini E. 2010. Rare variant association analysis methods for complex traits. *Annu. Rev. Genet.* 44:293–308

Asimit JL, Day-Williams AG, Morris AP, Zeggini E. 2012b. ARIEL and AMELIA: testing for an accumulation of rare variants using next-generation sequencing data. *Hum. Hered.* 73:84–94
Aulchenko YS, de Koning DJ, Haley C. 2007. Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. *Genetics* 177:577–85
Ayers KL, Lange K. 2008. Penalized estimation of haplotype frequencies. *Bioinformatics* 24:1596–602
Bacanu SA, Nelson MR, Whittaker JC. 2012. Comparison of statistical tests for association between rare variants and binary traits. *PLoS ONE* 7:e42530
Bansal V, Libiger O, Torkamani A, Schork NJ. 2010. Statistical analysis strategies for association studies involving rare variants. *Nat. Rev. Genet.* 11:773–85
Beck A, Teboulle M. 2009. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.* 2:183–202
Behar DM, Yunusbayev B, Metspalu M, Metspalu E, Rosset S, et al. 2010. The genome-wide structure of the Jewish people. *Nature* 466:238–42
Blangero J, Diego VP, Dyer TD, Almeida M, Peralta J, et al. 2013. A kernel of truth: statistical advances in polygenic variance component models for complex human pedigrees. *Adv. Genet.* 81:1–31
Bodmer W, Bonilla C. 2008. Common and rare variants in multifactorial susceptibility to common diseases. *Nat. Genet.* 40:695–701
Bodmer WF. 2010. Commentary: connections between genetics and statistics: a commentary on Fisher's 1951 Bateson lecture—'Statistical Methods in Genetics.' *Int. J. Epidemiol.* 39:340–44
Boehnke M, Cox NJ. 1997. Accurate inference of relationships in sib-pair linkage studies. *Am. J. Hum. Genet.* 61:423–29
Böhning D, Lindsay BG. 1988. Monotonicity of quadratic-approximation algorithms. *Ann. Inst. Stat. Math.* 40:641–63
Borwein JM, Lewis AS. 2006. *Convex Analysis and Nonlinear Optimization: Theory and Examples*. New York: Springer. 2nd ed.
Breheny P, Huang J. 2011. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Ann. Appl. Stat.* 5:232–53
Broman KW, Sen S. 2009. *A Guide to QTL Mapping with R/qtl*. New York: Springer
Browning SR, Browning BL. 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* 81:1084–97
Cai JF, Candès E, Shen Z. 2010. A singular value thresholding algorithm for matrix completion. *SIAM J. Optim.* 20:1956–82
Candès EJ, Tao T. 2010. The power of convex relaxation: near-optimal matrix completion. *IEEE Trans. Inf. Theory* 56:2053–80
Cantor RM, Lange K, Sinsheimer JS. 2010. Prioritizing GWAS results: a review of statistical methods and recommendations for their application. *Am. J. Hum. Genet.* 86:6–22
Chen C, He B, Yuan X. 2012a. Matrix completion via an alternating direction method. *IMA J. Numer. Anal.* 32:227–45
Chen GK, Wang K, Stram AH, Sobel EM, Lange K. 2012b. Mendel-GPU: haplotyping and genotype imputation on graphics processing units. *Bioinformatics* 28:2979–80
Chen SS, Donoho DL, Saunders MA. 1998. Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.* 20:33–61
Chi EC, Zhou H, Chen GK, Del Vecchyo DO, Lange K. 2013. Genotype imputation via matrix completion. *Genome Res.* 23:509–18
Claerbout J, Muir F. 1973. Robust modeling with erratic data. *Geophysics* 38:826–44
Day-Williams AG, Blangero J, Dyer TD, Lange K, Sobel EM. 2011. Linkage analysis without defined pedigrees. *Genet. Epidemiol.* 35:360–70
de Leeuw J. 1994. Block-relaxation algorithms in statistics. In *Information Systems and Data Analysis: Prospects, Foundations, Applications*, ed. HH Bock, W Lenski, MM Richter, pp. 308–24. Berlin: Springer
Derkach A, Lawless JF, Sun L. 2013. Robust and powerful tests for rare variants using Fisher's method to combine evidence of association from two or more complementary tests. *Genet. Epidemiol.* 37:110–21

Dziuda DM. 2010. *Data Mining for Genomics and Proteomics: Analysis of Gene and Protein Expression Data*. New York: Wiley. 1st ed.
Efron B. 2010. *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. New York: Cambridge Univ. Press
Elston RC, Stewart J. 1971. A general model for the genetic analysis of pedigree data. *Hum. Hered.* 21:523–42
Excoffier L, Slatkin M. 1995. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol. Biol. Evol.* 12:921–27
Friedman J, Hastie T, Höfling H, Tibshirani R. 2007. Pathwise coordinate optimization. *Ann. Appl. Stat.* 1:302–32
Friedman J, Hastie T, Tibshirani R. 2010. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* 33:1–22
Gibson G. 2012. Rare and common variants: twenty arguments. *Nat. Rev. Genet.* 13:135–45
Haseman JK, Elston RC. 1972. The investigation of linkage between a quantitative trait and a marker locus. *Behav. Genet.* 2:3–19
Henn BM, Gignoux CR, Jobin M, Granka JM, Macpherson JM, et al. 2011. Hunter-gatherer genomic diversity suggests a southern African origin for modern humans. *Proc. Natl. Acad. Sci. USA* 108:5154–62
Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, et al. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. USA* 106:9362–67
Hoffman GE, Logsdon BA, Mezey JG. 2013. PUMA: a unified framework for penalized multiple regression analysis of GWAS data. *PLoS Comput. Biol.* 9:e1003101
Hopper JL, Mathews JD. 1982. Extensions to multivariate normal models for pedigree analysis. *Ann. Hum. Genet.* 46:373–83
Horvath S. 2011. *Weighted Network Analysis: Applications in Genomics and Systems Biology*. New York: Springer
Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. 2012. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet.* 44:955–59
Hunter DR, Lange K. 2004. A tutorial on MM algorithms. *Am. Stat.* 58:30–37
Int. HapMap Consort. 2003. The International HapMap Project. *Nature* 426:789–96
Ionita-Laza I, Buxbaum JD, Laird NM, Lange C. 2011. A new testing strategy to identify rare variants with either risk or protective effect on disease. *PLoS Genet.* 7:e1001289
Jacquard A. 1970. *Structures Génétiques des Populations.* Paris: Masson Cie
Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, et al. 2010. Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* 42:348–54
Kiezun A, Garimella K, Do R, Stitziel NO, Neale BM, et al. 2012. Exome sequencing and the genetic basis of complex traits. *Nat. Genet.* 44:623–30
Kijas JW, Lenstra JA, Hayes B, Boitard S, Porto Neto LR, et al. 2012. Genome-wide analysis of the world's sheep breeds reveals high levels of historic mixture and strong recent selection. *PLoS Biol.* 10:e1001258
Ku CS, Loy EY, Pawitan Y, Chia KS. 2010. The pursuit of genome-wide association studies: Where are we now? *J. Hum. Genet.* 55:195–206
Kumar R, Seibold MA, Aldrich MC, Williams LK, Reiner AP, et al. 2010. Genetic ancestry in lung-function predictions. *N. Engl. J. Med.* 363:321–30
Laird NM, Lange C. 2011. *The Fundamentals of Modern Statistical Genetics*. New York: Springer
Lange K. 2002. *Mathematical and Statistical Methods for Genetic Analysis*. New York: Springer. 2nd ed.
Lange K. 2010. *Numerical Analysis for Statisticians*. New York: Springer
Lange K, Papp JC, Sinsheimer JS, Sripracha R, Zhou H, Sobel EM. 2013. Mendel: the Swiss army knife of genetic analysis programs. *Bioinformatics* 29:1568–70
Latorre V, Diskin SJ, Diamond MA, Zhang H, Hakonarson H, et al. 2012. Replication of neuroblastoma SNP association at the *BARD1* locus in African-Americans. *Cancer Epidemiol. Biomark. Prev.* 21:658–63
Lee SH, Wray NR, Goddard ME, Visscher PM. 2011. Estimating missing heritability for disease from genome-wide association studies. *Am. J. Hum. Genet.* 88:294–305
Li B, Leal SM. 2008. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am. J. Hum. Genet.* 83:311–21

Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. 2010. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.* 34:816–34

Liu DJ, Leal SM. 2012. A unified framework for detecting rare variant quantitative trait associations in pedigree and unrelated individuals via sequence data. *Hum. Hered.* 73:105–22

Long JC, Williams RC, Urbanek M. 1995. An E-M algorithm and testing strategy for multiple-locus haplotypes. *Am. J. Hum. Genet.* 56:799–810

Madsen BE, Browning SR. 2009. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet.* 5:e1000384

Mazumder R, Friedman JH, Hastie T. 2011. SparseNet: coordinate descent with nonconvex penalties. *J. Am. Stat. Assoc.* 106:1125–38

Mazumder R, Hastie T, Tibshirani R. 2010. Spectral regularization algorithms for learning large incomplete matrices. *J. Mach. Learn. Res.* 11:2287–322

McLachlan GJ, Krishnan T. 2007. *The EM Algorithm and Extensions*. Hoboken, NJ: Wiley. 2nd ed.

Mechanic LE, Chen HS, Amos CI, Chatterjee N, Cox NJ, et al. 2012. Next generation analytic tools for large scale genetic epidemiology studies of complex diseases. *Genet. Epidemiol.* 36:22–35

Meier L, van de Geer S, Bühlmann P. 2008. The group lasso for logistic regression. *J. R. Stat. Soc. B* 70:53–71

Meinshausen N, Bühlmann P. 2010. Stability selection. *J. R. Stat. Soc. B* 72:417–73

Morgenthaler S, Thilly WG. 2007. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutat. Res.* 615:28–56

Morris GP, Ramu P, Deshpande SP, Hash CT, Shah T, et al. 2013. Population genomic and genome-wide association studies of agroclimatic traits in sorghum. *Proc. Natl. Acad. Sci. USA* 110:453–58

Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, et al. 2011. Testing for an unusual distribution of rare variants. *PLoS Genet.* 7:e1001322

Nesterov Y. 2007. *Gradient methods for minimizing composite objective function.* CORE Discuss. Pap., Univ. Cathol. Louvain, Belg. **http://www.ucl.be/cps/ucl/doc/core/documents/coredp2007_76.pdf**

Pagani L, Kivisild T, Tarekegn A, Ekong R, Plaster C, et al. 2012. Ethiopian genetic diversity reveals linguistic stratification and complex influences on the Ethiopian gene pool. *Am. J. Hum. Genet.* 91:83–96

Pasaniuc B, Rohland N, McLaren PJ, Garimella K, Zaitlen N, et al. 2012. Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. *Nat. Genet.* 44:631–35

Price AL, Kryukov GV, de Bakker PIW, Purcell SM, Staples J, et al. 2010. Pooled association tests for rare variants in exon-resequencing studies. *Am. J. Hum. Genet.* 86:832–38

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* 38:904–9

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. *Genetics* 155:945–59

Rasmussen M, Li Y, Lindgreen S, Pedersen JS, Albrechtsen A, et al. 2010. Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature* 463:757–62

Risch N, Merikangas K. 1996. The future of genetic studies of complex human diseases. *Science* 273:1516–17

Sampson J, Jacobs K, Yeager M, Chanock S, Chatterjee N. 2011. Efficient study design for next generation sequencing. *Genet. Epidemiol.* 35:269–77

Sánchez E, Rasmussen A, Riba L, Acevedo-Vasquez E, Kelly JA, et al. 2012. Impact of genetic ancestry and sociodemographic status on the clinical expression of systemic lupus erythematosus in American Indian–European populations. *Arthritis Rheum.* 64:3687–94

Santosa F, Symes W. 1986. Linear inversion of band-limited reflection seismograms. *SIAM J. Sci. Stat. Comput.* 7:1307–30

Scheet P, Stephens M. 2006. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.* 78:629–44

Shah SC, Kusiak A. 2004. Data mining and genetic algorithm based gene/SNP selection. *Artif. Intell. Med.* 31:183–96

Singh AP, Pe'er I, Zafer S. 2013. MetaSeq: privacy preserving meta-analysis of sequencing-based association studies. *Pac. Symp. Biocomput.* 2013:356–67

Smith CAB. 1957. Counting methods in genetical statistics. *Ann. Hum. Genet.* 21:254–76

Stephens M, Smith NJ, Donnelly P. 2001. A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.* 68:978–89

Storey JD, Tibshirani R. 2003. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. USA* 100:9440–45

Strachan T, Read AP. 2011. *Human Molecular Genetics*. New York: Garland Sci./Taylor & Francis Group. 4th ed.

Taylor H, Banks S, McCoy J. 1979. Deconvolution with the ℓ_1 norm. *Geophysics* 44:39–52

Thomas DC. 2004. *Statistical Methods in Genetic Epidemiology*. New York: Oxford Univ. Press

Thompson EA. 1974. Gene identities and multiple relationships. *Biometrics* 30:667–80

Thompson EA. 1975. The estimation of pairwise relationships. *Ann. Hum. Genet.* 39:173–88

Tibshirani R. 1996. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B* 58:267–88

Tibshirani R, Saunders M, Rosset S, Zhu J, Knight K. 2005. Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc. B* 67:91–108

van Heel DA, Franke L, Hunt KA, Gwilliam R, Zhernakova A, et al. 2007. A genome-wide association study for celiac disease identifies risk variants in the region harboring *IL2* and *IL21*. *Nat. Genet.* 39:827–29

Visscher PM, Brown MA, McCarthy MI, Yang J. 2012. Five years of GWAS discovery. *Am. J. Hum. Genet.* 90:7–24

Williams AL, Patterson N, Glessner J, Hakonarson H, Reich D. 2012. Phasing of many thousands of genotyped samples. *Am. J. Hum. Genet.* 91:238–51

Williams JT, Blangero J. 1999. Power of variance component linkage analysis to detect quantitative trait loci. *Ann. Hum. Genet.* 63:545–63

Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. 2011. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* 89:82–93

Wu TT, Chen YF, Hastie T, Sobel E, Lange K. 2009. Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics* 25:714–21

Wu TT, Lange K. 2008. Coordinate descent algorithms for lasso penalized regression. *Ann. Appl. Stat.* 2:224–44

Wu TT, Lange K. 2010. The MM alternative to EM. *Stat. Sci.* 25:492–505

Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, et al. 2010. Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* 42:565–69

Yang J, Manolio TA, Pasquale LR, Boerwinkle E, Caporaso N, et al. 2011. Genome partitioning of genetic variation for complex traits using common SNPs. *Nat. Genet.* 43:519–25

Yuan M, Lin Y. 2006. Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. B* 68:49–67

Zhang T. 2010. Analysis of multi-stage convex relaxation for sparse regularization. *J. Mach. Learn. Res.* 11:1081–107

Zhang Z, Lange K, Ophoff R, Sabatti C. 2010. Reconstructing DNA copy number by penalized estimation and imputation. *Ann. Appl. Stat.* 4:1749–73

Zhang Z, Lange K, Sabatti C. 2012. Reconstructing DNA copy number by joint segmentation of multiple sequences. *BMC Bioinforma.* 13:205

Zhou H, Alexander D, Lange K. 2011a. A quasi-Newton acceleration for high-dimensional optimization algorithms. *Stat. Comput.* 21:261–73

Zhou H, Alexander DH, Sehl ME, Sinsheimer JS, Sobel EM, Lange K. 2011b. Penalized regression for genome-wide association screening of sequence data. *Pac. Symp. Biocomput.* 2011:106–17

Zhou H, Sehl ME, Sinsheimer JS, Lange K. 2010. Association screening of common and rare genetic variants by penalized regression. *Bioinformatics* 26:2375–82

Zhou JJ, Ghazalpour A, Sobel EM, Sinsheimer JS, Lange K. 2012. Quantitative trait loci association mapping by imputation of strain origins in multifounder crosses. *Genetics* 190:459–73

Ziegler A, König IR, Pahlke F. 2010. *A Statistical Approach to Genetic Epidemiology: Concepts and Applications*. Weinheim, Ger.: Wiley-VCH. 2nd ed.

Annual Review of Statistics and Its Application

Volume 1, 2014

Contents