Fast gradient-based optimization of resistance surfaces

Nathaniel S. Pope*†

Key point for introduction: this is a flexible framework for resistance surface optimization, that is fast because (A) it only requires a single Cholesky decomposition; (B) it uses analytic gradients with little additional computational cost.

1 Isolation by resistance as a non-linear model

The fundamental idea behind isolation by resistance is that a landscape may be discretized into a fully-connected graph (e.g. a lattice), and the movement of organisms modelled as a random walk on this graph. One of the simplest models choices of random walk is a symmetric, continuous-time Markov process: random walkers at the ith vertex of the graph transition to an adjacent vertex j at rate λ_{ij} (where $\lambda_{ij} = \lambda_{ji}$). Competing hypotheses about how landscape heterogeneity impacts the movement of organisms may be directly encoded by particular choices of transition rates ("conductances") among graph vertices: thus this framework has had great utility in landscape ecology since its introduction by McRae (2005).

A measure of connectivity that arises naturally from this framework is the commute distance (aka "resistance distance") between two locations: this is the expected time taken by a random walker to move from one point to another point B and back again. The infintesimal generator of the Markov process is the graph Laplacian $\mathbf{Q}$ with entries

$$\mathbf{Q}_{ij} = \begin{cases} -\lambda_{ij} & \text{if } i \neq j, j \in \mathcal{N}_i \\ \sum_k \lambda_{ik} & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

where $\mathcal{N}_i$ is the set of vertices that are neighbours to vertex i. $\mathbf{Q}$ is thus a very sparse, symmetric, positive semidefinite matrix of rank $N - 1$, where N is the number of vertices in the graph. Let $\mathbf{Z}$ be an N by M indicator matrix mapping a set of M spatial locations to vertices, so that $\mathbf{Z}_{ik} = 1$ if the kth spatial location is at vertex i, and 0 otherwise. The matrix of resistance distances between the spatial locations encoded by $\mathbf{Z}$ is

$$\mathbf{R} = \operatorname{diag}\{\mathbf{Z}^T\mathbf{Q}^+\mathbf{Z}\}\mathbf{1}^T + \mathbf{1}\operatorname{diag}\{\mathbf{Z}^T\mathbf{Q}^+\mathbf{Z}\}^T - 2\mathbf{Z}^T\mathbf{Q}^+\mathbf{Z} \tag{1}$$

*Department of Entomology, the Pennsylvania State University, University Park, PA 16802
†nsp5229@psu.edu

where $\mathbf{Q}^+$ is the generalized inverse of $\mathbf{Q}$. Ideally, resistance distances would be compared to observed migration rates, and a strong correspondence between the two would be taken as evidence for the underlying hypothesis of landscape conductance to movement. In practice, migration is difficult to measure directly, and genetic divergence among spatially-indexed individuals is used as a proxy.

For the remainder of this paper, I model the conductance of an edge connecting vertices i and j as separable and additive, so that $\lambda_{ij} = \mathbf{C}_i + \mathbf{C}_j$. For the ith vertex, I model the contribution to conductance as a function of a vector of covariates $\mathbf{x}_i$, so that $\mathbf{C}_i = f(\mathbf{x}_i; \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a vector of unknown parameters and $\mathbf{C}_i > 0$. For example, a log-linear model for vertex conductance is $f(\mathbf{x}_i; \boldsymbol{\theta}) = \exp\{\mathbf{x}_i^T \boldsymbol{\theta}\}$ (an intercept is omitted as it simply rescales $\mathbf{R}$). I refer to f as the "conductance model", and to the vector $\mathbf{C}(\boldsymbol{\theta}) = [\mathbf{C}_1, \ldots, \mathbf{C}_N]$ as the "conductance surface" generated by $\boldsymbol{\theta}$.

Let $\mathbf{Y}$ be either the genotypes of individuals collected across the locations encoded by $\mathbf{Z}$, or a function of these genotypes such as a genetic dissimilarity metric. Let $g(\mathbf{Y}; \mathbf{R})$ be the probability of $\mathbf{Y}$ given resistance distances $\mathbf{R}$. For example, g might be the likelihood of a linear regression between genetic relatedness and resistance distance (I discuss possible choices of g in detail below). I refer to g as the "measurement model". The log-likelihood of $\boldsymbol{\theta}$ given the genetic data and spatial covariates is $\mathcal{L}(\boldsymbol{\theta}) = \log g(\mathbf{Y}; \mathbf{R}(\mathbf{Q}(\mathbf{C}(\boldsymbol{\theta}))))$. The maximum likelihood estimate $\hat{\boldsymbol{\theta}} = \arg\max \mathcal{L}(\boldsymbol{\theta})$ provides an estimate of the conductance surface $\mathbf{C}(\hat{\boldsymbol{\theta}})$ for a given set of spatial covariates.

Model selection – that is, choosing among different sets of spatial covariates or different functional forms for f – can proceed via standard practices for maximum likelihood inference, such as likelihood ratio tests, cross-validation, or information criteria, using the number of free values in θ as the degrees of freedom. For example, the log-likelihood under the null model of isolation by distance is $\mathcal{L}_{\text{IBD}} = \log g(\mathbf{Y}; \mathbf{R}(\mathbf{Q}(\mathbf{C}(\mathbf{0}))))$ where $\mathbf{C}(\mathbf{0}) = [1, \ldots, 1]$. For the log-linear model $\mathbf{C}_i = \exp\{\mathbf{x}_i^T \boldsymbol{\theta}\}$ (and for many other reasonable choices of f), the null model is on the interior of the parameter space (e.g. the point where $\boldsymbol{\theta} = 0$), so $2\mathcal{L}(\hat{\boldsymbol{\theta}}) - 2\mathcal{L}_{\text{IBD}}$ is asymptotically distributed as $\chi^2(\|\hat{\boldsymbol{\theta}}\|_0)$.

Many prior studies (REF) have ... This is not ideal: ... However, the framework described above is not novel: several prior works have also considered optimization in frameworks that are implicitly or explicitly similar to the one described above. The challenge – and the novel contribution of this study – is to do so in a computationally efficient manner.

2 Efficient gradient calculation

Provided that $g(\mathbf{Y}; \mathbf{R})$ and $f(\boldsymbol{\theta})$ are differentiable in $\mathbf{R}$ and $\boldsymbol{\theta}$ respectively, first-order derivative information can greatly accelarate optimization of $\mathcal{L}$, especially when $\boldsymbol{\theta}$ has many elements. However, numeric approximation by Richardson extrapolation or finite differencing is costly, because the large linear system in Equation 1 must be solved for

every perturbation of the parameter vector. Here, I use reverse algorithmic differentiation (repeated application of the chain rule in reverse) to give an exact algorithm for the gradient $\dot{\boldsymbol{\theta}} = [\partial\mathcal{L}/\partial\boldsymbol{\theta}_1, \ldots, \partial\mathcal{L}/\partial\boldsymbol{\theta}_K]$ that requires virtually no computational cost beyond that already used to calculate $\mathcal{L}$.

The graph Laplacian $\mathbf{Q}$ has a single zero eigenvalue, and its nullspace is spanned by the vector $\mathbf{v} = N^{-1/2}\mathbf{1}$ [**?**]. Let the rank $N-1$ diagonalization of $\mathbf{Q} = \mathbf{P}\mathbf{D}\mathbf{P}^T$ (e.g. the columns of $\mathbf{P}$ are the $N-1$ eigenvectors spanning the column space). As the column space and nullspace are orthogonal, $\mathbf{P}\mathbf{P}^T + \mathbf{v}\mathbf{v}^T = \mathbf{I}$. Let $\mathbf{I}_{-N}$ be the N-dimensional identity matrix with column N removed. Then,

$$\begin{aligned}\mathbf{Q}^+ = \mathbf{P}\mathbf{D}^{-1}\mathbf{P}^T &= \mathbf{P}\mathbf{P}^T\mathbf{I}_{-N}^T(\mathbf{P}^T\mathbf{I}_{-N}^T)^{-1}\mathbf{D}^{-1}(\mathbf{I}_{-N}\mathbf{P})^{-1}\mathbf{I}_{-N}\mathbf{P}\mathbf{P}^T \\ &= (\mathbf{I}_{-N}^T - \mathbf{v}\mathbf{v}^T\mathbf{I}_{-N}^T)(\mathbf{P}^T\mathbf{I}_{-N}^T)^{-1}\mathbf{D}^{-1}(\mathbf{I}_{-N}\mathbf{P})^{-1}(\mathbf{I}_{-N} - \mathbf{v}\mathbf{v}^T\mathbf{I}_{-N}) \\ &= (\mathbf{I}_{-N}^T - \mathbf{v}\mathbf{v}^T\mathbf{I}_{-N}^T)(\mathbf{I}_{-N}\mathbf{Q}\mathbf{I}_{-N}^T)^{-1}(\mathbf{I}_{-N} - \mathbf{v}\mathbf{v}^T\mathbf{I}_{-N}).\end{aligned}$$

From a computational perspective, this identity is very useful, because the problem in Equation 1 reduces to solving $\mathbf{Q}_{-N}\mathbf{G} = \mathbf{Z}_{-N}$ for $\mathbf{G}$, where $\mathbf{Q}_{-N} = \mathbf{I}_{-N}\mathbf{Q}\mathbf{I}_{-N}^T$ and $\mathbf{Z}_{-N} = (\mathbf{I}_{-N} - \mathbf{v}\mathbf{v}^T\mathbf{I}_{-N})\mathbf{Z}$, so that

$$\mathbf{R} = \operatorname{diag}\{\mathbf{Z}_{-N}^T\mathbf{G}\}\mathbf{1}^T + \mathbf{1}\operatorname{diag}\{\mathbf{Z}_{-N}^T\mathbf{G}\}^T - 2\mathbf{Z}_{-N}^T\mathbf{G}.$$

This requires a single Cholesky decomposition of the very sparse, full-rank matrix $\mathbf{Q}_{-N}$ (algorithm **??**A). Other algorithms in use (e.g. [**?**]) require the solution of M independent linear systems (equivalently M Cholesky decompositions). The identity derived above is closely related to the method described by [**?**].

Let $\dot{\mathbf{R}}$ be the symmetric matrix of partial derivatives of $\log g(\mathbf{Y};\mathbf{R})$ with respect to resistance distances $\mathbf{R}$. This is typically cheap to compute (specific examples are given below). Using the chain rule in reverse, the gradient with respect to the i,jth element of $\mathbf{Q}_{-N}$ is:

$$(\dot{\mathbf{Q}}_{-N})_{ij} = \begin{cases}\mathbf{G}_i(\mathbf{I}\circ\dot{\mathbf{R}}\mathbf{1}\mathbf{1}^T - \dot{\mathbf{R}})\mathbf{G}_i^T & \text{if } i = j \\ -\mathbf{G}_i(\mathbf{I}\circ\dot{\mathbf{R}}\mathbf{1}\mathbf{1}^T - \dot{\mathbf{R}})\mathbf{G}_j^T & \text{if } i \neq j, j \in \mathcal{N}_i \\ 0 & \text{otherwise,}\end{cases} \tag{2}$$

where $\mathbf{G}_i$ is the ith row of $\mathbf{G}$ and $\circ$ is the element-wise product. The gradient with respect to the elements in the conductance surface is

$$\dot{\mathbf{C}}_i = \begin{cases} n_i[\dot{\mathbf{Q}}_{-N}]_{ii} + \sum_{j\in\mathcal{N}_i\setminus N}[\dot{\mathbf{Q}}_{-N}]_{jj} - 2[\dot{\mathbf{Q}}_{-N}]_{ij} & \text{if } i \neq N \\ \sum_{j\in\mathcal{N}_N}[\dot{\mathbf{Q}}_{-N}]_{jj} & \text{if } i = N,\end{cases}$$

where $n_i = \operatorname{card}\{\mathcal{N}_i\}$ is the number of neighbours of vertex i. This scheme is very computationally efficient as it involves iterating over only the non-zero elements of $\mathbf{Q}$ and utilizes the previously computed $\mathbf{G}$ (algorithm **??**B). Finally, $\dot{\boldsymbol{\theta}}_k = \sum_i \dot{\mathbf{C}}_i(\partial\mathbf{C}_i/\partial\boldsymbol{\theta}_k)$. For example, $\partial\mathbf{C}_i/\partial\boldsymbol{\theta}_k = \mathbf{x}_{ki}\mathbf{C}_i$ for the log-linear model $\mathbf{C}_i = \exp\{\mathbf{x}_i^T\boldsymbol{\theta}\}$. Importantly, this calculation is exact and the computational cost is essentially independent of the dimension of $\boldsymbol{\theta}$, unlike gradient approximation by finite differencing.

3 Measurement models

Here I derive the gradient $\dot{\mathbf{R}}$ required in the previous section for four reasonable choices of $g(\mathbf{Y};\mathbf{R})$. Let $\mathbf{S}$ be a genetic dissimilarity matrix calculated from genotypes $\mathbf{Y}$. Let $\mathbf{s},\mathbf{r}$ be the lower triangular vectorizations of $\mathbf{S},\mathbf{R}$ (e.g. the $\binom{M}{2}$ values of $\mathbf{R}_{ij}$ where $i<j$). For clarity, I retain matrix indexing despite vectorization: e.g. $\mathbf{r}_{ij}=\mathbf{R}_{ij}$.

Naive regression. The model is $\mathbf{s}=\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\epsilon}$, where $\mathbf{X}=[\mathbf{1},\mathbf{r}]$ and $\boldsymbol{\epsilon}$ are *i.i.d.* Gaussian errors with standard deviation σ. Given $\mathbf{r}$, the maximum likelihood estimates for the nuisance parameters are $\hat{\boldsymbol{\beta}}=(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{s}$ and $\hat{\sigma}^2=\binom{M}{2}^{-1}\hat{\mathbf{e}}^T\hat{\mathbf{e}}$ where $\hat{\mathbf{e}}=\mathbf{s}-\mathbf{X}\hat{\boldsymbol{\beta}}$. Thus $\log g_{\text{reg}}(\mathbf{Y};\mathbf{R})=-2^{-1}\hat{\sigma}^{-2}\hat{\mathbf{e}}^T\hat{\mathbf{e}}-\binom{M}{2}\log\hat{\sigma}$, and

$$[\dot{\mathbf{R}}_{\text{reg}}]_{ij}=[\dot{\mathbf{R}}_{\text{reg}}]_{ji}=2^{-1}\hat{\sigma}^{-2}\hat{\boldsymbol{\beta}}_2\hat{\mathbf{e}}_{ij}.$$

Maximum likelihood population effects. Naive regression between $\mathbf{s}$ and $\mathbf{r}$ assumes that the errors are independent, which is unlikely to hold as $\mathbf{s}$ contains pairwise measurements. The maximum likelihood population effects (MLPE) model of [**?**] attempts to model the dependence between pairwise observations by introducing *i.i.d.* Gaussian effects $\boldsymbol{\gamma}$, one for each spatial location: $\mathbf{s}_{ij}=[\mathbf{X}\boldsymbol{\beta}]_{ij}+\boldsymbol{\gamma}_i+\boldsymbol{\gamma}_j+\boldsymbol{\epsilon}_{ij}$. Let $\mathbf{U}$ be an $\binom{M}{2}$-by-M indicator matrix mapping spatial locations to pairwise observations, so that $\mathbf{U}_{ij,k}=1$ if $k\in\{i,j\}$ and 0 otherwise. After integrating over $\boldsymbol{\gamma}$, the correlation of $\mathbf{s}$ becomes $\boldsymbol{\Sigma}=(1-2\rho)\mathbf{I}+\rho\mathbf{U}\mathbf{U}^T$ and the maximum likelihood estimates are $\hat{\boldsymbol{\beta}}=(\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{s}$ and $\hat{\sigma}=\binom{M}{2}^{-1}\hat{\mathbf{e}}^T\boldsymbol{\Sigma}^{-1}\hat{\mathbf{e}}$ where $\hat{\mathbf{e}}=\mathbf{s}-\mathbf{X}\hat{\boldsymbol{\beta}}$. Thus $\log g_{\text{mlpe}}(\mathbf{Y};\mathbf{R},\rho)=-2^{-1}\hat{\sigma}^{-2}\hat{\mathbf{e}}^T\boldsymbol{\Sigma}^{-1}\hat{\mathbf{e}}-2^{-1}\binom{M}{2}\log\det(\sigma^2\boldsymbol{\Sigma})$, and

$$[\dot{\mathbf{R}}_{\text{mlpe}}]_{ij}=[\dot{\mathbf{R}}_{\text{mlpe}}]_{ji}=2^{-1}\hat{\sigma}^{-2}\hat{\boldsymbol{\beta}}_2[\boldsymbol{\Sigma}^{-1}\hat{\mathbf{e}}]_{ij}.$$

The nuisance parameter ρ has no closed form optimum and must be optimized along with $\boldsymbol{\theta}$; note that $\partial\mathcal{L}/\partial\rho=\text{tr}\{(\mathbf{I}-\sigma^{-2}\boldsymbol{\Sigma}^{-1}\hat{\mathbf{e}}\hat{\mathbf{e}}^T)\boldsymbol{\Sigma}^{-1}(\mathbf{I}-2^{-1}\mathbf{U}^T\mathbf{U})\}$. Any matrix-vector product involving $\boldsymbol{\Sigma}^{-1}$ may be efficiently computed via the Woodbury lemma [**?**].

Generalized Wishart. [**?**] suggest using $\mathbf{E}=\mathbf{Z}^T\mathbf{Q}^+\mathbf{Z}$ itself as a covariance structure for the normalized genotypes. For simplicity, assume the genotypes are biallelic across L loci – although what follows can be extended to multiallelic markers – so that $\mathbf{Y}$ and $\mathbf{N}$ are L-by-M matrices of derived allele counts and number of sampled haplotypes, respectively. For locus l, let the global allele frequencies $\mathbf{F}_l=(\sum_i\mathbf{N}_{li})^{-1}\sum_i\mathbf{Y}_{li}$ and normalized genotypes $\tilde{\mathbf{Y}}_{li}=(\mathbf{N}_{li}\mathbf{F}_l-\mathbf{N}_{li}\mathbf{F}_l^2)^{-1/2}(\mathbf{Y}_{li}-\mathbf{N}_{li}\mathbf{F}_l)$. Let the genetic dissimilarity matrix $\mathbf{S}=\mathbf{1}\,\text{diag}\{\tilde{\mathbf{Y}}^T\tilde{\mathbf{Y}}\}^T+\text{diag}\{\tilde{\mathbf{Y}}^T\tilde{\mathbf{Y}}\}\mathbf{1}^T-2\tilde{\mathbf{Y}}^T\tilde{\mathbf{Y}}$. If the columns of $\tilde{\mathbf{Y}}^T$ are *i.i.d.* multivariate Gaussian vectors with covariance $\boldsymbol{\Sigma}=\tau^2\mathbf{E}+\sigma^2\mathbf{I}$, then $\mathbf{S}$ follows a generalized Wishart distribution [**?**], and $\log g_{\text{wis}}(\mathbf{Y};\mathbf{E},\sigma^2)=4^{-1}L\,\text{tr}\{\boldsymbol{\Sigma}^{-1}\mathbf{W}\mathbf{S}\}+2^{-1}L\log\text{Det}\{\boldsymbol{\Sigma}^{-1}\mathbf{W}\}$, where Det is the pseudo-determinant (the product of non-zero eigenvalues) and $\mathbf{W}=\mathbf{I}-\mathbf{1}(\mathbf{1}^T\boldsymbol{\Sigma}^{-1}\mathbf{1})^{-1}\mathbf{1}^T\boldsymbol{\Sigma}^{-1}$ is a projection onto the column space of $\boldsymbol{\Sigma}^{-1}$.

Because $\mathbf{E}$ is used in place of $\mathbf{R}$, for the calculation of $\dot{\boldsymbol{\theta}}$ Equation 2 should be replaced by

$$(\dot{\mathbf{Q}}_{-N})_{ij} = \begin{cases} \mathbf{G}_i \dot{\mathbf{E}} \mathbf{G}_i^T & \text{if } i = j \\ -\mathbf{G}_i \dot{\mathbf{E}} \mathbf{G}_j^T & \text{if } i \neq j, j \in \mathcal{N}_i \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

Using a result from [**?**] for differentiation of the pseudo-determinant,

$$\dot{\mathbf{E}}_{\text{wis}} = -L\tau^2 \boldsymbol{\Sigma}^{-1} \mathbf{W} (2^{-1} (\boldsymbol{\Sigma}^{-1} \mathbf{W})^+ + 4^{-1} \mathbf{S}) \mathbf{W}^T \boldsymbol{\Sigma}^{-1}$$

The nuisance parameters σ^2 and τ^2 must be optimized in tandem with $\boldsymbol{\theta}$; note that $\partial \mathcal{L} / \partial \tau^2 = \tau^{-2} \operatorname{tr}\{\mathbf{E} \dot{\mathbf{E}}_{\text{wis}}\}$ and $\partial \mathcal{L} / \partial \sigma^2 = \tau^{-2} \operatorname{tr}\{\dot{\mathbf{E}}_{\text{wis}}\}$. In practice, dissimilarity measures such as F_{st} could be substituted for the distance matrix $\mathbf{S}$ defined above, although the rationale behind the derivation would be lost.

Binomial mixed model. In some cases, the normalized genotypes $\tilde{\mathbf{Y}}$ may not be well approximated by a Gaussian, for example when there are only a few sampled haplotypes per spatial location (e.g. a single diploid individual). It may be more appropriate to model the raw genotypes $\mathbf{Y}$ as binomial. Let $\mathbf{z}_{il} = \alpha + \beta_l + \boldsymbol{\epsilon}_{il}$ be a latent variable, and let $\mathbf{p}_{il} = (1 + e^{-\mathbf{z}_{il}})^{-1}$ be the probability that a randomly sampled haplotype from location i carries the derived allele at locus l. The parameter α is the average allele frequency across loci, the per-locus effects β_l are *i.i.d.* Gaussian with standard deviation ω, and the errors $\boldsymbol{\epsilon}_l$ are Gaussian vectors with covariance $\tau^2 \mathbf{E} + \sigma^2 \mathbf{I}$. Thus $\mathbf{z}_l$ are Gaussian vectors with covariance $\boldsymbol{\Sigma} = \tau^2 \mathbf{E} + \omega^2 \mathbf{1}\mathbf{1}^T + \sigma^2 \mathbf{I}$ and mean $\boldsymbol{\mu} = \alpha \mathbf{1}$. The likelihood is intractable, but can be reasonably approximated by Laplace's method [**?**]. Fisher scoring is used to find the maximum $\hat{\mathbf{z}}_l$ of $h(\mathbf{z}_l) = -2^{-1} (\mathbf{z}_l - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{z}_l - \boldsymbol{\mu}) + \sum_i \mathbf{Y}_{il} \log \mathbf{p}_{il} + (\mathbf{N}_{il} - \mathbf{Y}_{il}) \log(1 - \mathbf{p}_{il})$ (Appendix **??**). Then, $\log g_{\text{bin}}(\mathbf{Y}; \mathbf{E}, \alpha, \sigma^2, \tau^2, \omega^2) = 2^{-1} (L \log\det \boldsymbol{\Sigma}^{-1} - \sum_l h(\hat{\mathbf{z}}_l) - \log\det \mathbf{H}_{\hat{\mathbf{z}}_l})$ where $[\mathbf{H}_{\hat{\mathbf{z}}_l}]_{ij} = [\boldsymbol{\Sigma}^{-1}]_{ij} - \mathbb{I}[i = j] \mathbf{N}_{il} \hat{\mathbf{p}}_{il} (1 - \hat{\mathbf{p}}_{il})$. The adjoint method (Appendix **??**) can be used to derive the gradient,

$$\dot{\mathbf{E}}_{\text{bin}} = ???,$$

where ... The nuisance parameters have gradients $\partial \mathcal{L} / \partial \alpha = ???, \partial \mathcal{L} / \partial \sigma^2 = ???, \partial \mathcal{L} / \partial \tau^2 = ..., \partial \mathcal{L} / \partial \omega^2 = ???$.

Both of the regression-based measurement models (g_{reg} and g_{mlpe}) allow an unconstrained linear relationship between $\mathbf{r}$ and $\mathbf{s}$. Thus, for these choices of g, local optima of $\mathcal{L}(\boldsymbol{\theta})$ may exist where genetic distance *decreases* with resistance distance (e.g. $\hat{\boldsymbol{\beta}}_2 < 0$). As this is implausible on biological grounds, the model can be modified to incorporate the constraint $\boldsymbol{\beta}_2 > 0$. For the log-linear conductance $\mathbf{C}_i = \exp\{x_i^T \boldsymbol{\theta}\}$, this is accomplished by absorbing $\boldsymbol{\beta}_2$ into $\boldsymbol{\theta}$ (Appendix **??**).

4 Hessian and leverage

For both optimization and calculation of asymptotic standard errors, it is useful to have an efficient calculation of the Hessian.

Consider the second partial derivatives of the likelihood with respect to elements of $\mathbf{Q}_{-N}$. In other words:

$$\partial[\partial\mathcal{L}/[\mathbf{Q}_{-N}]_{ij}]/\partial[\mathbf{Q}_{-N}]_{kl} = \partial[\dot{\mathbf{Q}}_{-N}]_{ij}/\partial[\mathbf{Q}_{-N}]_{kl}$$

summing over i, j and using standard matrix calculus,

$$\sum_{i,j} \partial[\dot{\mathbf{Q}}_{-N}]_{ij}/\partial[\mathbf{Q}_{-N}]_{kl} = -[\mathbf{Q}_{-N}^{-1}\mathbf{1}\mathbf{1}^T\mathbf{Q}_{-N}^{-1}\mathbf{Z}\dot{\mathbf{E}}\mathbf{Z}^T\mathbf{Q}^{-1}]_{kl}$$

That's the second derivative of the likelihood with regard to elements in Q_{-N}.

Have $\mathcal{L}(\hat{\theta}(x_i))$ where x_i are the spatial covariates, and $\hat{\theta}(x_i)$ emphasizes that the MLE is a function of the spatial covariates. We have $dL/dtheta \equiv 0$. Thus $d(dL/dtheta)/dx = 0$. And $d(dL/dtheta)/dx = d2L/dtheta2dtheta/dx + dL/dxdtheta$. We know dL/dx from above.

Finally, for the three methods with differentiable responses, can calculate leverage with regard to a measure of genetic distance.

5 Implementation and benchmarking

Numerical verification for the gradients derived in this paper is provided as a supplementary R script.

I recorded the time taken to compute likelihood and gradient versus likelihood alone, and the number of iterations needed to find $\hat{\boldsymbol{\theta}}$ using gradient-free (BOBQYA) and gradient-based (BFGS) optimization algorithms, across varying numbers of sampled spatial locations and spatial covariates. As a point of comparison for my implementation, I also used Circuitscape [version information, settings].

Simulate data from coalescent to test asymptotic bias, because measurement models used above are all drastic simplifications of actual geneological process.

6 Example: gene flow in *M. ???*