Appendix 5

Derivatives of Vectors and Vector-Valued Functions

Draft Version 24 November 2008

Quantitative genetics often deals with vector-valued functions and here we provide a brief review of the calculus of such functions. In particular, we review common expressions for derivatives of vectors and vector-valued functions, introduce the gradient vector and Hessian matrix (for first and second partials, respectively), and then use this machinery in multidimensional Taylor series for approximating functions around a specific value. We apply these results to several problems in selection theory and evolution.

DERIVATIVES OF VECTORS AND VECTOR-VALUED FUNCTIONS

Let $f(\mathbf{z})$ be a scaler (single dimensional) function of a vector $\mathbf{x}$ of n variables, $x_1, \cdots, x_n$. The **gradient** (or **gradient vector**) of f with respect to $\mathbf{x}$ is obtained by taking partial derivatives of the function with respect to each variable. In matrix notation, the gradient operator is

$$\nabla_{\mathbf{x}}[\,f\,] = \frac{\partial f}{\partial \mathbf{x}} = \begin{pmatrix} \dfrac{\partial f}{\partial x_1} \\ \dfrac{\partial f}{\partial x_2} \\ \vdots \\ \dfrac{\partial f}{\partial x_n} \end{pmatrix}$$

The gradient at point $\mathbf{x}_o$ corresponds to a vector indicating the direction of local steepest ascent of the function at that point (the multivariate slope of f at the point $\mathbf{x}_o$).

Example A5.1. Compute the gradient for

$$f(\mathbf{x}) = \sum_{i=1}^{n} x_i^2 = \mathbf{x}^T\mathbf{x}$$

Since $\partial f / \partial x_i = 2x_i$, the gradient vector is just $\nabla_{\mathbf{x}}[\,f(\mathbf{x})\,] = 2\mathbf{x}$. At the point $\mathbf{x}_o$, $\mathbf{x}^T\mathbf{x}$ locally increases most rapidly if we change $\mathbf{x}$ in the same the direction as the vector going from point $\mathbf{x}_o$ to point $\mathbf{x}_o + 2\delta\,\mathbf{x}_o$, where δ is a small positive value.

For a vector $\mathbf{a}$ and matrix $\mathbf{A}$ of constants, it can easily be shown (e.g., Morrison 1976, Graham 1981, Searle 1982) that

$$\nabla_{\mathbf{x}} \left[\mathbf{a}^T \mathbf{x} \right] = \nabla_{\mathbf{x}} \left[\mathbf{x}^T \mathbf{a} \right] = \mathbf{a} \tag{A5.1a}$$

$$\nabla_{\mathbf{x}} \left[\mathbf{A} \mathbf{x} \right] = \mathbf{A}^T \tag{A5.1b}$$

Turning to quadratic forms, if $\mathbf{A}$ is symmetric, then

$$\nabla_{\mathbf{x}} \left[\mathbf{x}^T \mathbf{A} \mathbf{x} \right] = 2 \cdot \mathbf{A} \mathbf{x} \tag{A5.1c}$$

$$\nabla_{\mathbf{x}} \left[(\mathbf{x} - \mathbf{a})^T \mathbf{A} (\mathbf{x} - \mathbf{a}) \right] = 2 \cdot \mathbf{A} (\mathbf{x} - \mathbf{A}) \tag{A5.1d}$$

$$\nabla_{\mathbf{x}} \left[(\mathbf{a} - \mathbf{x})^T \mathbf{A} (\mathbf{a} - \mathbf{x}) \right] = -2 \cdot \mathbf{A} (\mathbf{a} - \mathbf{x}) \tag{A5.1e}$$

Taking $\mathbf{A} = \mathbf{I}$, Equation A5.1c implies

$$\nabla_{\mathbf{x}} \left[\mathbf{x}^T \mathbf{x} \right] = \nabla_{\mathbf{x}} \left[\mathbf{x}^T \, \mathbf{I} \, \mathbf{x} \right] = 2 \cdot \mathbf{I} \, \mathbf{x} = 2 \cdot \mathbf{x} \tag{A5.1f}$$

as was found in Example A5.1. Two other useful identities follow from the chain rule of differentiation,

$$\nabla_{\mathbf{x}} \left[\exp[\, f(\mathbf{x}) \,] \right] = \exp[\, f(\mathbf{x}) \,] \cdot \nabla_{\mathbf{x}} \left[f(\mathbf{x}) \right] \tag{A5.1g}$$

$$\nabla_{\mathbf{x}} \left[\ln[\, f(\mathbf{x}) \,] \right] = \frac{1}{f(\mathbf{x})} \cdot \nabla_{\mathbf{x}} \left[f(\mathbf{x}) \right] \tag{A5.1h}$$

Finally, the product rule also applies to a gradient, with

$$\nabla_{\mathbf{x}} \left[f(\mathbf{x}) \, g(\mathbf{x}) \,] \right] = \nabla_{\mathbf{x}} \left[f(\mathbf{x}) \,] \right] \, g(\mathbf{x}) + f(\mathbf{x}) \, \nabla_{\mathbf{x}} \left[g(\mathbf{x}) \,] \right] \tag{A5.1i}$$

Example A5.2 The multivariate normal (MVN) distribution returns a scaler value and is a function of the data vector $\mathbf{x}$, the vector of means $\boldsymbol{\mu}$, and the covariance matrix $\mathbf{V_x}$ We can write the multivariate normal (MVN) distribution function as

$$\varphi(\mathbf{x}) = a \exp \left(-\frac{1}{2} \cdot (\mathbf{x} - \boldsymbol{\mu})^T \, \mathbf{V_x}^{-1} \, (\mathbf{x} - \boldsymbol{\mu}) \right)$$

where the constant $a = \pi^{-n/2} \, |\mathbf{V_x}|^{-1/2}$. The compute the gradient of the MNV with respect to the data vector $\mathbf{x}$, first apply Equation A5.1g,

$$\nabla_{\mathbf{x}} \left[\varphi(\mathbf{x}) \right] = \varphi(\mathbf{x}) \cdot \nabla_{\mathbf{x}} \left[\left(-\frac{1}{2} \right) \cdot (\mathbf{x} - \boldsymbol{\mu})^T \, \mathbf{V_x}^{-1} \, (\mathbf{x} - \boldsymbol{\mu}) \right]$$

Applying Equation A5.1d gives

$$\nabla_{\mathbf{x}} \left[\varphi(\mathbf{x}) \right] = -\varphi(\mathbf{x}) \cdot \mathbf{V_x}^{-1} \, (\mathbf{x} - \boldsymbol{\mu}) \tag{A5.2a}$$

Note here that $\varphi(\mathbf{x})$ is a scalar and hence its order of multiplication does not matter, while the order of the other variables (being matrices) is critical. Similarly, Equation A5.1e implies the gradient of the MVN with respect to the vector of means $\boldsymbol{\mu}$ is

$$\nabla_{\boldsymbol{\mu}} \left[\varphi(\mathbf{x}, \boldsymbol{\mu}) \right] = \varphi(\mathbf{x}, \boldsymbol{\mu}) \cdot \mathbf{V_x}^{-1} \, (\mathbf{x} - \boldsymbol{\mu}) \tag{A5.2b}$$

Example A5.3. Recall (Chapters 10, 16, 29) that the directional selection gradient $\boldsymbol{\beta} = \mathbf{P}^{-1}\mathbf{S}$. The reason that $\boldsymbol{\beta}$ is called a gradient is because Lande (1979) showed that $\boldsymbol{\beta}$ equals the gradient of log mean fitness with respect to the vector of trait means, $\nabla_{\boldsymbol{\mu}}[\,\ln \overline{W}(\boldsymbol{\mu})\,]$, when phenotypes are multivariate normal. Hence, the increase in mean population fitness is maximized if mean character values change in the same direction as the vector $\boldsymbol{\beta}$. To see this, first note that applying Equation A5.1h gives

$$\nabla_{\boldsymbol{\mu}}[\,\ln \overline{W}(\boldsymbol{\mu})\,] = \overline{W}^{\,-1}\,\nabla_{\boldsymbol{\mu}}[\,\overline{W}(\boldsymbol{\mu})\,]$$

Writing mean fitness as $\overline{W}(\boldsymbol{\mu}) = \int W(\mathbf{z})\,\varphi(\mathbf{z},\boldsymbol{\mu})\,d\mathbf{z}$ and taking the gradient through the integral gives

$$\nabla_{\boldsymbol{\mu}}[\,\ln \overline{W}(\boldsymbol{\mu})\,] = \nabla_{\boldsymbol{\mu}}\left[\int \frac{W(\mathbf{z})}{\overline{W}}\,\varphi(\mathbf{z},\boldsymbol{\mu})\,\mathrm{d}\mathbf{z}\right] = \int w(\mathbf{z})\,\nabla_{\boldsymbol{\mu}}\,[\,\varphi(\mathbf{z},\boldsymbol{\mu})\,]\;\;\mathrm{d}\mathbf{z}$$

If individual fitnesses are frequency-dependent (functions of the population mean instead of fixed constants), then a second integral appears as we can no longer assume $\nabla_{\boldsymbol{\mu}}\,[\,w(\mathbf{z})\,] = 0$ (see Equation 29.5b). If the trait vector $\mathbf{z}$ is distributed MVN($\boldsymbol{\mu}, \mathbf{P}$), the Equation A5.2b gives

$$\nabla_{\boldsymbol{\mu}}\,[\,\varphi(\mathbf{z},\boldsymbol{\mu})\,] = \mathbf{P}^{\,-1}\,(\mathbf{z}-\boldsymbol{\mu})$$

Hence we can rewrite the above integral as

$$\begin{aligned}\int w(\mathbf{z})\,\nabla_{\boldsymbol{\mu}}\,[\,\varphi(\mathbf{z},\boldsymbol{\mu})\,]\;\mathrm{d}\mathbf{z} &= \int w(\mathbf{z})\,\varphi(\mathbf{z})\,\mathbf{P}^{\,-1}\,(\mathbf{z}-\boldsymbol{\mu})\,\mathrm{d}\mathbf{z} \\ &= \mathbf{P}^{\,-1}\left(\int \mathbf{z}\,w(\mathbf{z})\,\varphi(\mathbf{z})\,\mathrm{d}\mathbf{z} - \boldsymbol{\mu}\int w(\mathbf{z})\,\varphi(\mathbf{z})\,\mathrm{d}\mathbf{z}\right) \\ &= \mathbf{P}^{\,-1}(\boldsymbol{\mu}^{*}-\boldsymbol{\mu}) = \mathbf{P}^{\,-1}\mathbf{S} = \boldsymbol{\beta}\end{aligned}$$

which follows since the first integral (in the second above line) is the mean character value after selection and the second equals one as $E[\,w\,] = 1$ by definition.

Example A5.4. A common class of fitness functions used in models of the evolution of quantitative traitsis the **Gaussian fitness function**,

$$\begin{aligned}W(\mathbf{z}) &= \exp\left(\boldsymbol{\alpha}^T\mathbf{z} - \frac{1}{2}(\mathbf{z}-\boldsymbol{\theta})^T\mathbf{W}(\mathbf{z}-\boldsymbol{\theta})\right) \\ &= \exp\left(\sum_i \alpha_i z_i - \frac{1}{2}\sum_i\sum_j (z_i-\theta_i)(z_j-\theta_j)W_{ij}\right)\end{aligned}$$

where $\mathbf{W}$ is a symmetric matrix. If a vector of traits is distributed as a MVN($\boldsymbol{\mu}, \mathbf{P}$) before selection, then after selection the distribution remains multivariate normal with new mean and variance

$$\boldsymbol{\mu}^{*} = \mathbf{P}^{*}(\mathbf{P}^{-1}\boldsymbol{\mu} + \mathbf{W}\boldsymbol{\theta} + \boldsymbol{\alpha})$$

where

$$\mathbf{P}^{*} = \mathbf{P} - \mathbf{P}\left(\mathbf{P}+\mathbf{W}^{-1}\right)^{-1}\mathbf{P}$$

A fair bit of work (**Example XX.xx**) shows that the vector of selection differentials can be expressed as

$$\mathbf{S} = \boldsymbol{\mu}^* - \boldsymbol{\mu} = \mathbf{W}^{-1}\left(\mathbf{W}^{-1} + \mathbf{P}\right)^{-1}\mathbf{P}\left(\mathbf{W}(\boldsymbol{\theta} - \boldsymbol{\mu}) + \boldsymbol{\alpha}\right) \tag{A5.3a}$$

The resulting population mean fitness is

$$\overline{W}(\boldsymbol{\mu}, \mathbf{P}) = \sqrt{\frac{|\,\mathbf{P}^*\,|}{|\,\mathbf{P}\,|}} \cdot \exp\left(-\frac{1}{2}\left[f(\boldsymbol{\mu})\right]\right)$$

where

$$f(\boldsymbol{\mu}) = \boldsymbol{\theta}^T\mathbf{W}\boldsymbol{\theta} + \boldsymbol{\mu}^T\mathbf{P}^{-1}\left(\mathbf{I} - \mathbf{P}^*\mathbf{P}^{-1}\right)\boldsymbol{\mu} - 2\cdot\mathbf{b}^T\mathbf{P}^{-1}\boldsymbol{\mu} - \mathbf{b}^T\mathbf{P}^{-1}\mathbf{b}$$

with $\mathbf{b} = \mathbf{W}\boldsymbol{\theta} + \boldsymbol{\alpha}$. Let's compute the gradient in log mean fitness with respect to the vector of means. Taking the log of fitness gives

$$\nabla_{\boldsymbol{\mu}}\left[\,\ln\overline{W}(\boldsymbol{\mu}, \mathbf{P})\,\right] = \nabla_{\boldsymbol{\mu}}\left[\ln\left(\frac{|\,\mathbf{P}^*\,|}{|\,\mathbf{P}\,|}\right)\right] - \frac{1}{2}\cdot\nabla_{\boldsymbol{\mu}}\left[\,f(\boldsymbol{\mu})\,\right] = -\frac{1}{2}\cdot\nabla_{\boldsymbol{\mu}}\left[\,f(\boldsymbol{\mu})\,\right]$$

where the first term is zero because $\mathbf{P}$ and $\mathbf{P}^*$ are independent of $\boldsymbol{\mu}$. Ignoring terms of f not containing $\boldsymbol{\mu}$ since the gradient of these (with respect to $\boldsymbol{\mu}$) is zero,

$$\nabla_{\boldsymbol{\mu}}\left[\,f(\boldsymbol{\mu})\,\right] = \nabla_{\boldsymbol{\mu}}\left[\,\boldsymbol{\mu}^T\mathbf{P}^{-1}\left(\mathbf{I} - \mathbf{P}^*\mathbf{P}^{-1}\right)\boldsymbol{\mu}\,\right] - 2\cdot\nabla_{\boldsymbol{\mu}}\left[\mathbf{b}^T\mathbf{P}^{-1}\boldsymbol{\mu}\right]$$

Applying Equations A5.1b/c,

$$\nabla_{\boldsymbol{\mu}}\left[\,\boldsymbol{\mu}^T\mathbf{P}^{-1}\left(\mathbf{I} - \mathbf{P}^*\mathbf{P}^{-1}\right)\boldsymbol{\mu}\,\right] = 2\cdot\mathbf{P}^{-1}\left(\mathbf{I} - \mathbf{P}^*\mathbf{P}^{-1}\right)\boldsymbol{\mu}$$
$$\nabla_{\boldsymbol{\mu}}\left[\mathbf{b}^T\mathbf{P}^{-1}\boldsymbol{\mu}\right] = \left(\mathbf{b}^T\mathbf{P}^{-1}\right)^T = \mathbf{P}^{-1}\mathbf{b}$$

Hence,

$$\nabla_{\boldsymbol{\mu}}\left[\,\ln\overline{W}(\boldsymbol{\mu}, \mathbf{P})\,\right] = \mathbf{P}^{-1}\left[\,\left(\mathbf{P}^*\mathbf{P}^{-1} - \mathbf{I}\right)\boldsymbol{\mu} + \mathbf{b}\,\right] \tag{A5.3b}$$

Using the definitions of $\mathbf{P}^*$ and $\mathbf{b}$, we can (eventually) express this as

$$\nabla_{\boldsymbol{\mu}}\left[\,\ln\overline{W}(\boldsymbol{\mu}, \mathbf{P})\,\right] = \mathbf{P}^{-1}\mathbf{W}^{-1}\left(\mathbf{W}^{-1} + \mathbf{P}\right)^{-1}\mathbf{P}\left(\mathbf{W}(\boldsymbol{\theta} - \boldsymbol{\mu}) + \boldsymbol{\alpha}\right) \tag{A5.3c}$$

Recalling Equation A5.3a, this is just $\mathbf{P}^{-1}\mathbf{S} = \boldsymbol{\beta}$, as expected.

Example A5.5. Consider obtaining the least-squares solution for the general linear model, $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, where we wish to find the value of $\boldsymbol{\beta}$ that minimizes the residual error given $\mathbf{y}$ and $\mathbf{X}$. In matrix form,

$$\begin{aligned}\sum_{i=1}^{n} e_i^2 &= \mathbf{e}^T\mathbf{e} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{x}\boldsymbol{\beta}) \\ &= \mathbf{y}^T\mathbf{y} - \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} \\ &= \mathbf{y}^T\mathbf{y} - 2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}\end{aligned}$$

where the last step follows since the matrix product $\boldsymbol{\beta}^T \mathbf{X}^T \mathbf{y}$ yields a scaler, and hence it equals its transpose,

$$\boldsymbol{\beta}^T \mathbf{X}^T \mathbf{y} = \left(\boldsymbol{\beta}^T \mathbf{X}^T \mathbf{y}\right)^T = \mathbf{y}^T \mathbf{X} \boldsymbol{\beta}$$

To find the vector $\boldsymbol{\beta}$ that minimizes $\mathbf{e}^T \mathbf{e}$, taking the derivative with respect to $\boldsymbol{\beta}$ and using Equations A5.1a/c gives

$$\frac{\partial\, \mathbf{e}^T \mathbf{e}}{\partial\, \boldsymbol{\beta}} = -2\mathbf{X}^T \mathbf{y} + 2\mathbf{X}^T \mathbf{X} \boldsymbol{\beta}$$

Setting this equal to zero gives $\mathbf{X}^T \mathbf{X} \boldsymbol{\beta} = \mathbf{X}^T \mathbf{y}$ giving

$$\boldsymbol{\beta} = \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{y}$$

More generally, if $\mathbf{X}^T \mathbf{X}$ is singular, we can still solve this equation by using a generalized inverse $\left(\mathbf{X}^T \mathbf{X}\right)^{-}$, see LW Appendix 3.

The Hessian Matrix, Local Maxima/minima, and Multidimensional Taylor Series

In univariate calculus, local extrema of a function occur when the slope (first derivative) is zero. The multivariate extension is that the gradient vector is zero, so that the slope of the function with respect to all variables is zero. A point $\mathbf{x}_e$ where this occurs is called a **stationary** or **equilibrium** point, and corresponds to either a local maximum, minimum, saddle point or inflection point. As with the calculus of single variables, determining which of these is true depends on the second derivative. With n variables, the appropriate generalization is the **hessian** matrix

$$\mathbf{H}_{\mathbf{x}}[\, f\,] = \nabla_{\mathbf{x}} \left[\left(\nabla_{\mathbf{x}} [\, f\,] \right)^T \right] = \frac{\partial^2 f}{\partial \mathbf{x}\, \partial \mathbf{x}^T} = \begin{pmatrix} \dfrac{\partial^2 f}{\partial x_1^2} & \cdots & \dfrac{\partial^2 f}{\partial x_1\, \partial x_n} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial^2 f}{\partial x_1\, \partial x_n} & \cdots & \dfrac{\partial^2 f}{\partial x_n^2} \end{pmatrix} \tag{A5.4}$$

This matrix is symmetric, as mixed partials are equal under suitable continuity conditions, and measures the local multidimensional curvature of the function.

Example A5.6. Compute $\mathbf{H}_{\mathbf{x}}\,[\,\varphi(\mathbf{x})\,]$, the hessian matrix for the multivariate normal distribution. Recalling from Equation A5.2a that $\nabla_{\mathbf{x}}\,[\,\varphi(\mathbf{x})\,] = -\varphi(\mathbf{x}) \cdot \mathbf{V}_{\mathbf{X}}^{-1}\,(\mathbf{x} - \boldsymbol{\mu})$, we have

$$\begin{aligned} \mathbf{H}_{\mathbf{x}}\,[\,\varphi(\mathbf{x})\,] &= \nabla_{\mathbf{x}} \left[\left(\nabla_{\mathbf{x}}\,[\,\varphi(\mathbf{x})\,] \right)^T \right] \\ &= -\nabla_{\mathbf{x}}\left[\,\varphi(\mathbf{x}) \cdot (\mathbf{x} - \boldsymbol{\mu})^T \mathbf{V}_{\mathbf{X}}^{-1}\,\right] \\ &= -\nabla_{\mathbf{x}}\,[\,\varphi(\mathbf{x})\,] \cdot (\mathbf{x} - \boldsymbol{\mu})^T \mathbf{V}_{\mathbf{X}}^{-1} - \varphi(\mathbf{x}) \cdot \nabla_{\mathbf{x}}\left[\,(\mathbf{x} - \boldsymbol{\mu})^T \mathbf{V}_{\mathbf{X}}^{-1}\,\right] \\ &= \varphi(\mathbf{x}) \cdot \left(\mathbf{V}_{\mathbf{X}}^{-1}\,(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T \mathbf{V}_{\mathbf{X}}^{-1} - \mathbf{V}_{\mathbf{X}}^{-1} \right) \end{aligned} \tag{A5.5a}$$

Where we have used Equations A5.1i and A5.1b, respectively (recall that $\mathbf{V_X}$ is a vector of constants). In a similiar manner, the gradient with respect to the vector of means is

$$\mathbf{H}_{\boldsymbol{\mu}}\left[\,\varphi(\mathbf{x},\boldsymbol{\mu})\,\right] = \varphi(\mathbf{x},\boldsymbol{\mu})\cdot\left(\,\mathbf{V}_{\mathbf{X}}^{-1}\,(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{V}_{\mathbf{X}}^{-1} - \mathbf{V}_{\mathbf{X}}^{-1}\right) \tag{A5.5b}$$

To see how the hessian matrix determines the nature of equilibrium points, a slight digression on the multidimensional Taylor series is needed. Consider the (second-order) Taylor series of a function of n variables $f(x_1,\cdots,x_n)$ expanded about the point $\mathbf{y}$,

$$f(\mathbf{x}) \simeq f(\mathbf{y}) + \sum_{i=1}^{n}(x_i-y_i)\frac{\partial f}{\partial x_i} + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i-y_i)(x_j-y_j)\frac{\partial^2 f}{\partial x_i\,\partial x_j} + \cdots$$

where all partials are evaluated at $\mathbf{y}$. Noting that the first sum is just the inner product of the gradient and $\mathbf{x}-\mathbf{x}_o$ while the double sum is a quadratic product involving the Hessian, we can express this in matrix form as

$$f(\mathbf{x}) \simeq f(\mathbf{x}_o) + \nabla^T(\mathbf{x}-\mathbf{x}_o) + \frac{1}{2}(\mathbf{x}-\mathbf{x}_o)^T\mathbf{H}\,(\mathbf{x}-\mathbf{x}_o) \tag{A5.6}$$

where ∇ and $\mathbf{H}$ are the gradient and hessian with respect to $\mathbf{x}$ evaluated at $\mathbf{x}_o$, e.g.,

$$\nabla \equiv \nabla_{\mathbf{x}}[\,f\,]\,\big|_{\mathbf{x}=\mathbf{x}_o} \qquad \text{and} \qquad \mathbf{H} \equiv \mathbf{H}_{\mathbf{x}}[\,f\,]\,\big|_{\mathbf{x}=\mathbf{x}_o}$$

Example A5.7. Consider the following demonstration (due to Lande 1979) that mean population fitness increases. A round of selection changes the current vector of means from $\boldsymbol{\mu}$ to $\boldsymbol{\mu}+\boldsymbol{\Delta\mu}$. Expanding the log of mean fitness in a Taylor series around the current population mean $\boldsymbol{\mu}$ gives the change in mean population fitness as

$$\begin{aligned}\Delta\ln\overline{W}(\boldsymbol{\mu}) &= \ln\overline{W}(\boldsymbol{\mu}+\boldsymbol{\Delta\mu}) - \ln\overline{W}(\boldsymbol{\mu}) \\ &\simeq \left(\nabla_{\boldsymbol{\mu}}[\,\ln\overline{W}(\boldsymbol{\mu})\,]\right)^T\boldsymbol{\Delta\mu} + \frac{1}{2}\boldsymbol{\Delta\mu}^T\,\mathbf{H}_{\boldsymbol{\mu}}[\,\ln\overline{W}(\boldsymbol{\mu})\,]\,\boldsymbol{\Delta\mu}\end{aligned}$$

assuming that second and higher-order terms can be neglected (as would occur with weak selection and the population mean away from an equilibrium point), then

$$\Delta\ln\overline{W}(\boldsymbol{\mu}) \simeq \left(\nabla_{\boldsymbol{\mu}}[\,\ln\overline{W}(\boldsymbol{\mu})\,]\right)^T\boldsymbol{\Delta\mu}$$

Assuming that the joint distribution of phenotypes and additive genetic values is MVN, then $\boldsymbol{\Delta\mu}=\mathbf{G}\boldsymbol{\beta}$, or $\nabla_{\boldsymbol{\mu}}[\,\ln\overline{W}(\boldsymbol{\mu})\,]=\boldsymbol{\beta}=\mathbf{G}^{-1}\boldsymbol{\Delta\mu}$. Substituting gives

$$\Delta\ln\overline{W}(\boldsymbol{\mu}) \simeq \left(\mathbf{G}^{-1}\boldsymbol{\Delta\mu}\right)^T\boldsymbol{\Delta\mu} = (\boldsymbol{\Delta\mu})^T\,\mathbf{G}^{-1}\,\boldsymbol{\Delta\mu} \geq 0$$

The inequality follows since $\mathbf{G}$ is a variance-covariance matrix and hence is non-negative definite (see below). Thus under these conditions, mean population fitness always increases, although since $\boldsymbol{\Delta\mu}\neq\nabla_{\boldsymbol{\mu}}[\,\ln\overline{W}(\boldsymbol{\mu})\,]$ fitness does not increase in the fastest possible manner.

At an equilibrium point $\widehat{\mathbf{x}}$, all first partials are zero, so that $(\nabla_{\mathbf{x}}[\,f\,])^T$ at this point is a vector of length zero. Whether this point is a maximum or minimum is then determined by the quadratic product involving the hessian evaluated at $\widehat{\mathbf{x}}$. Considering vector $\mathbf{d}$ of a small change from the equilibrium point,

$$f(\widehat{\mathbf{x}}+\mathbf{d}) - f(\widehat{\mathbf{x}}) \simeq \frac{1}{2} \cdot \mathbf{d}^T\,\mathbf{H}\,\mathbf{d} \tag{A5.7a}$$

Since $\mathbf{H}$ is a symmetrix matrix, we can diagonalize it and apply a canonical transformation (Appendix 4) to simplify the quadratic form to give

$$f(\widehat{\mathbf{x}}+\mathbf{d}) - f(\widehat{\mathbf{x}}) \simeq \frac{1}{2} \sum_{i=1}^{n} \lambda_i y_i^2 \tag{A5.7b}$$

where $y_i = \mathbf{e}_i{}^T\mathbf{d}$, $\mathbf{e}_i$ and λ_i being the eigenvectors and eigenvalues of the hessian evaluated at $\widehat{\mathbf{x}}$. Thus, if $\mathbf{H}$ is **positive-definite** (all eigenvalues of $\mathbf{H}$ are positive), f increases in all directions around $\widehat{\mathbf{x}}$ and hence $\widehat{\mathbf{x}}$ is a local minimum of f. If $\mathbf{H}$ is **negative-definite** (all eigenvalues are negative), f decreases in all directions around $\widehat{\mathbf{x}}$ and $\widehat{\mathbf{x}}$ is a local maximum of f. If the eigenvalues differ in sign ($\mathbf{H}$ is **indefinite**), $\widehat{\mathbf{x}}$ corresponds to a saddle point (to see this, suppose $\lambda_1 > 0$ and $\lambda_2 < 0$; any change along the vector $\mathbf{e}_1$ results in an increases in f, while any change along $\mathbf{e}_2$ results in a decrease in f).

Example A5.8. Consider again the generalized Gaussian fitness function (Example A5.4). It can be shown (**Exampl xx.xx**) that $\boldsymbol{\beta} = \nabla_{\boldsymbol{\mu}}\left[\,\ln \overline{W}(\boldsymbol{\mu}, \mathbf{P})\,\right] = \mathbf{0}$ when $\widehat{\boldsymbol{\mu}} = \boldsymbol{\theta} + \mathbf{W}^{-1}\boldsymbol{\alpha}$. Is this a local maximum or a minimum? Recalling Equation A5.3b, we have

$$\begin{aligned}
\mathbf{H}_{\boldsymbol{\mu}}\left[\,\overline{W}(\boldsymbol{\mu})\,\right] &= \nabla_{\boldsymbol{\mu}}\left[\left(\nabla_{\boldsymbol{\mu}}\left[\,\ln \overline{W}(\boldsymbol{\mu})\,\right]\right)^T\right] \\
&= \nabla_{\boldsymbol{\mu}}\left[\left(\mathbf{P}^{-1}\left(\mathbf{P}^*\mathbf{P}^{-1} - \mathbf{I}\right)\boldsymbol{\mu} + \mathbf{P}^{-1}\mathbf{b}\right)^T\right] \\
&= \mathbf{P}^{-1}\left(\mathbf{P}^*\mathbf{P}^{-1} - \mathbf{I}\right)
\end{aligned}$$

Recall from Example A5.4 that $\mathbf{P}^* = \mathbf{P} - \mathbf{P}\left(\mathbf{P} + \mathbf{W}^{-1}\right)^{-1}\mathbf{P}$, giving

$$\begin{aligned}
\mathbf{H}_{\boldsymbol{\mu}}\left[\,\overline{W}(\boldsymbol{\mu})\,\right] &= \mathbf{P}^{-1}\left[\left(\mathbf{I} - \mathbf{P}\left(\mathbf{P} + \mathbf{W}^{-1}\right)^{-1}\right)\mathbf{P}\mathbf{P}^{-1} - \mathbf{I}\right] \\
&= \mathbf{P}^{-1}\left[\mathbf{I} - \mathbf{I} - \mathbf{P}\left(\mathbf{P} + \mathbf{W}^{-1}\right)^{-1}\right] \\
&= -\left(\mathbf{P} + \mathbf{W}^{-1}\right)^{-1}
\end{aligned} \tag{A5.8}$$

This result was obtained for the case of $\boldsymbol{\alpha} = \mathbf{0}$ by Lande (1979). Recall from Appendix 4 that if λ is an eigenvalue of $\mathbf{A}$, then $-\lambda^{-1}$ is an eigenvalue of $-\mathbf{A}^{-1}$. Thus if all eigenvalues of the matrix $\mathbf{J} = \left(\mathbf{P} + \mathbf{W}^{-1}\right)$ are positive ($\mathbf{J}$ is positive-definite), all eigenvalues of $\mathbf{H}_{\boldsymbol{\mu}}\left[\,\overline{W}(\boldsymbol{\mu})\,\right]$ are negative and hence $\widehat{\boldsymbol{\mu}}$ corresponds to a local maximum in the mean population fitness surface. If $\mathbf{J}$ is positive-definite, them for all non-trivial (i.e. positive lenght) vectors $\mathbf{x}$,

$$\mathbf{x}^T\,\mathbf{J}\,\mathbf{x} = \mathbf{x}^T\,\left(\mathbf{P} + \mathbf{W}^{-1}\right)\,\mathbf{x} = \mathbf{x}^T\,\mathbf{P}\,\mathbf{x} + \mathbf{x}^T\,\mathbf{W}^{-1}\,\mathbf{x} > 0$$

Letting λ_i and $\mathbf{e}_i$ be the ith eigenvalue and associated unit eigenvector for $\mathbf{P}$ and likewise γ_i and $\mathbf{f}_i$ be the eigenvalue and associated unit eigenvector for $\mathbf{W}$, then applying the canonical transformation of each matrix,

$$\mathbf{x}^T\,\mathbf{J}\,\mathbf{x} = \sum_{i=1}^{n} y_i^2 \lambda_i + \sum_{i=1}^{n} z_i^2\,\gamma_i^{-1}$$

where $y_i = {\mathbf{e}_i}^T\,\mathbf{x}$ and $z_i = \mathbf{f}_i^T\,\mathbf{x}$. Since all the eigenvalues of $\mathbf{P}$ are positive, $\mathbf{J}$ is positive-definite if all eigenvalues of $\mathbf{W}$ are positive (implying stabilizing selection on all characters). More generally, using the constraint that the phenotypic covariance matrix after selection $\mathbf{P}^* = (\mathbf{P}^{-1} + \mathbf{W})^{-1}$ must be positive definite, we can show that if γ_i corresponds to a negative eigenvalue of $\mathbf{W}$ (disruptive selection among the axis given by $\mathbf{f}_i$), then fitness is at a local minimum along this axis.

Optimization under constraints

Occasionally we wish to find the maximum or minimum of a function subject to a constaint. The solution is to use **Lagrange multipliers**: suppose we wish to find the extrema of $f(\mathbf{x})$ subject to the constraint $h(\mathbf{x}) = c$. Construct a new function g by considering

$$g(\mathbf{x}, \lambda) = f(\mathbf{x}) - \lambda(\,h(\mathbf{x}) - c\,)$$

Since $h(\mathbf{x}) - c = 0$, the extrema of $g(\mathbf{x}, \lambda)$ correspond to the extrema of $f(\mathbf{x})$ under the constraint. Local maxima and minima are obtained by solving the series of equations

$$\nabla_{\mathbf{x}}[\,g(\mathbf{x}, \lambda)\,] = \nabla_{\mathbf{x}}[\,f(\mathbf{x})\,] - \lambda \cdot \nabla_{\mathbf{x}}[\,h(\mathbf{x})\,] = \mathbf{0}$$
$$\frac{d\,g(\mathbf{x}, \lambda)}{d\lambda} = h(\mathbf{x}) - c = 0$$

Observe that the second equation is satisfied by the constraint.

Example A5.9. Consider a new (univariate) character, an **index** I, which is a linear combination of n characters $(z_1, z_2, \cdots, z_n)$,

$$I = \mathbf{b}^T\,\mathbf{z} = \sum_{k=1}^{n} b_k\,z_k$$

where we further impose $\mathbf{b}^T\mathbf{b} = 1$. Denote the directional selection differential of the new character I by S_I and observe that if $\mathbf{S}$ is the vector of directional selection differentials for $\mathbf{z}$, then $S_I = \mathbf{b}^T\mathbf{S}$. We wish to solve for $\mathbf{b}$ such that for a fixed amount of selection on I ($S_I = r$) we maximize the response of another linear combination of $\mathbf{z}$, $\mathbf{A}^T\boldsymbol{\mu} = \sum a_k\,\mu_k$. Assuming the conditions leading to the multivariate breeders' equation hold, the function to maximize is

$$f(\mathbf{b}) = \mathbf{a}^T\boldsymbol{\Delta\mu} = \mathbf{a}^T\mathbf{G}\mathbf{P}^{-1}\mathbf{S}$$

under the associated constraint function

$$g(\mathbf{b}) - c = \mathbf{b}^T\mathbf{b} - 1 = 0$$

Since $S_I = \mathbf{b}^T\mathbf{S}$ and we have the constraint $\mathbf{b}^T\mathbf{b} = 1$ take $\mathbf{S} = r \cdot \mathbf{b}$ so that $S_I = \mathbf{b}^T\mathbf{S} = r \cdot \mathbf{b}^T\mathbf{b} = r$. Taking derivatives gives

$$\nabla_{\mathbf{x}}[\, g(\mathbf{x}, \lambda)\,] = r \cdot \nabla_{\mathbf{x}}[\, \mathbf{a}^T\mathbf{G}\mathbf{P}^{-1}\mathbf{b}\,] - \lambda \cdot \nabla_{\mathbf{x}}[\, \mathbf{b}^T\mathbf{b}\,] = r \cdot \left(\mathbf{a}^T\mathbf{G}\mathbf{P}^{-1}\right)^T - (2\lambda) \cdot \mathbf{b}$$

which is equal to zero when

$$\mathbf{b} = (2\lambda/r) \cdot \mathbf{P}^{-1}\mathbf{G}\mathbf{a}$$

where we can solve for λ by using the constraint $\mathbf{b}^T\mathbf{b} = 1$. Thus

$$I = c \cdot \left(\mathbf{P}^{-1}\mathbf{G}\mathbf{a}\right)^T \mathbf{z} \tag{A5.9}$$

where the constant c depends on the desired strength of selection r. This is Smith's **optimal selection index** (Smith 1936, Hazel 1943) who obtained it by a very different approach. Index selection is the subject of Chapters 32-33.

Example A5.9. As an application of the above optimal selection index, suppose we wish to maximize the change in mean population fitness. Expanding mean population fitness in a Taylor series gives, to first order,

$$\begin{aligned}
\overline{W}(\boldsymbol{\mu} + \boldsymbol{\Delta\mu}) - \overline{W}(\boldsymbol{\mu}) &\simeq \nabla_{\boldsymbol{\mu}}[\, \overline{W}(\boldsymbol{\mu})\,]^T \, \boldsymbol{\Delta\mu} \\
&= \overline{W} \cdot \left(\overline{W}^{\,-1} \cdot \nabla_{\boldsymbol{\mu}}[\, \overline{W}(\boldsymbol{\mu})\,]^T \, \boldsymbol{\Delta\mu}\right) \\
&= \overline{W} \cdot \left(\nabla_{\boldsymbol{\mu}}[\, \ln \overline{W}(\boldsymbol{\mu})\,]^T \, \boldsymbol{\Delta\mu}\right) \\
&= \overline{W} \cdot \boldsymbol{\beta}^T \, \boldsymbol{\Delta\mu}
\end{aligned}$$

Thus, the linear combination we wish to maximize is $\boldsymbol{\beta}^T \boldsymbol{\mu}$, and from above in taking $\boldsymbol{\beta} = \mathbf{a}$ gives the selection index that maximizes the change in mean population fitness as

$$I = (\mathbf{P}^{-1}\, \mathbf{G}\boldsymbol{\beta})\mathbf{z}$$

Literature Cited

Graham, A. 1981. *Kronecker products and matrix calculus with applications*. Halsted Press, New York.

Hazel, L. N. 1943. The genetic basis for constructing selection indexes. *Genetics* 28: 476–490

Lande, R. 1979. Quantitative genetic analysis of multivariate evolution, applied to brain:body size allometry. *Evolution* 33: 402–416.

Morrison, D. F. 1976. *Multivariate statistical methods*. McGraw-Hill, New York.

Searle, S. S. 1982. *Matrix algebra useful for statistics*. John WIley and Sons, New York.

Smith, H. F. 1936. A discriminant function for plant selection. *Ann. Eugen.* 7: 240–250.