Chapter 4
Linear Marker and Genome-Wide Selection Indices

Abstract There are two main linear marker selection indices employed in marker-assisted selection (MAS) to predict the net genetic merit and to select individual candidates as parents for the next generation: the linear marker selection index (LMSI) and the genome-wide LMSI (GW-LMSI). Both indices maximize the selection response, the expected genetic gain per trait, and the correlation with the net genetic merit; however, applying the LMSI in plant or animal breeding requires genotyping the candidates for selection; performing a linear regression of phenotypic values on the coded values of the markers such that the selected markers are statistically linked to quantitative trait loci that explain most of the variability in the regression model; constructing the marker score, and combining the marker score with phenotypic information to predict and rank the net genetic merit of the candidates for selection. On the other hand, the GW-LMSI is a single-stage procedure that treats information at each individual marker as a separate trait. Thus, all marker information can be entered together with phenotypic information into the GW-LMSI, which is then used to predict the net genetic merit and select candidates. We describe the LMSI and GW-LMSI theory and show that both indices are direct applications of the linear phenotypic selection index theory to MAS. Using real and simulated data we validated the theory of both indices.

4.1 The Linear Marker Selection Index

4.1.1 Basic Conditions for Constructing the LMSI

In Chap. 2, Sect. 2.1, we indicated ten basic conditions for constructing a valid linear phenotypic selection index (LPSI). These ten conditions are also necessary for the linear marker selection index (LMSI); however, in addition to those conditions, the LMSI also requires the following conditions:

1. The markers and the quantitative trait loci (QTL) should be in linkage disequilibrium in the population under selection.
2. The QTL effects should be combined additively both within and between loci.

J. J. Céron-Rojas, J. Crossa, *Linear Selection Indices in Modern Plant Breeding*,
https://doi.org/10.1007/978-3-319-91223-3_4

3. The QTL should be in coupling mode, that is, one of the initial lines should have all the alleles that have a positive effect on the chromosome, and the other lines should have all the negative effects.
4. The traits of interest should be affected by a few QTL with large effects (and possibly a number of very small QTL effects) rather than many small QTL effects.
5. The heritability of the traits should be low.
6. Markers correlated with the traits of interest should be identified.

Under these conditions, the LMSI should be more efficient than the LPSI, at least in the first selection cycles (Whittaker 2003; Moreau et al. 2007).

4.1.2 The LMSI Parameters

Let $y_i = g_i + e_i$ be the ith trait ($i = 1, 2, \ldots, t$, $t =$ number of traits), where $e_i \sim N(0, \sigma^2_{e_i})$ is the residual with expectation equal to zero and variance value $\sigma^2_{e_i}$, and N stands for normal distribution. Assuming that the QTL effects combine additively both within and between loci, the ith unobservable genetic value g_i can be written as

$$g_i = \sum_{k=1}^{N_Q} \alpha_k q_k, \tag{4.1}$$

where α_k is the effect of the kth QTL, q_k is the number of favorable alleles at the kth QTL (2, 1 or 0), and N_Q is the number of QTL affecting the ith trait of interest.

If the QTL effect values are not observable, the g_i values in Eq. (4.1) are also not observable; however, we can use a linear combination of the markers linked to the QTL (s_i) that affect the ith trait to predict the g_i value as

$$s_i = \sum_{j=1}^{M} \theta_j x_j, \tag{4.2}$$

where s_i is a predictor of g_i, θ_j is the regression coefficient of the linear regression model, x_j is the coded value of the jth markers (e.g., 1, 0, and -1 for marker genotypes *AA*, *Aa* and *aa* respectively), and M is the number of selected markers linked to the QTL that affect the ith trait. Equation (4.2) is called the *marker score* (Lande and Thompson 1990; Whittaker 2003) and this is the main reason why the LMSI is not equal to the LPSI described in Chap. 2. The number of selected markers is only a subset of potential markers linked to QTL in the population under selection; thus, the s_i values should be lower than or equal to the g_i values. One way of estimating the s_i values is to perform a linear regression of phenotypic values on the coded values of the markers, select markers that are statistically linked to

quantitative trait loci that explain most of the variability in the regression model, and then obtain the estimated value of s_i ($\widehat{s}_i$) as the sum of the products of the QTL effects linked to markers and multiplied by the marker coded values associated with the ith trait. Some authors (e.g., Moreau et al. 2007) call $\widehat{s}_i$ the molecular score; in this book, we call s_i the marker score and $\widehat{s}_i$ the estimated marker score.

The objective of the LMSI is to predict the net genetic merit of each individual and select the individuals with the highest net genetic merit for further breeding. In the LMSI context, the net genetic merit can be written as

$$H = \mathbf{w}'\mathbf{g} + \mathbf{w}_2'\mathbf{s} = \begin{bmatrix}\mathbf{w}' & \mathbf{w}_2'\end{bmatrix}\begin{bmatrix}\mathbf{g}\\ \mathbf{s}\end{bmatrix} = \mathbf{a}'\mathbf{z}, \tag{4.3}$$

where $\mathbf{g}' = \begin{bmatrix} g_1 & \cdots & g_q \end{bmatrix}$ is the vector of breeding values; $\mathbf{w}' = \begin{bmatrix} w_1 & \cdots & w_t \end{bmatrix}$ is the vector of economic weights associated with $\mathbf{g}$; $\mathbf{w}_2' = \begin{bmatrix} 0_1 & \cdots & 0_t \end{bmatrix}$ is a null vector associated with the vector of marker scores $\mathbf{s}' = \begin{bmatrix} s_1 & \cdots & s_t \end{bmatrix}$; s_i is the ith marker score; $\mathbf{a}' = \begin{bmatrix}\mathbf{w}' & \mathbf{w}_2'\end{bmatrix}$ and $\mathbf{z} = \begin{bmatrix}\mathbf{g}' & \mathbf{s}'\end{bmatrix}$.

The information provided by the marker score can be used in breeding programs to increase the accuracy of predicting the net genetic merit of the individuals under selection. The LMSI combines the phenotypic and marker scores to predict H in each selection cycle and can be written as

$$I_M = \boldsymbol{\beta}_y'\mathbf{y} + \boldsymbol{\beta}_s'\mathbf{s} = \begin{bmatrix}\boldsymbol{\beta}_y' & \boldsymbol{\beta}_s'\end{bmatrix}\begin{bmatrix}\mathbf{y}\\ \mathbf{s}\end{bmatrix} = \boldsymbol{\beta}'\mathbf{t}, \tag{4.4}$$

where $\boldsymbol{\beta}_y'$ and $\boldsymbol{\beta}_s$ are vectors of phenotypic and marker score weights respectively; $\mathbf{y}' = \begin{bmatrix} y_1 & \cdots & y_t \end{bmatrix}$ is the vector of trait phenotypic values and $\mathbf{s}$ was defined in Eq. (4.3); $\boldsymbol{\beta}' = \begin{bmatrix}\boldsymbol{\beta}_y' & \boldsymbol{\beta}_s'\end{bmatrix}$ and $\mathbf{t}' = \begin{bmatrix}\mathbf{y}' & \mathbf{s}'\end{bmatrix}$.

The LMSI selection response can be written as

$$R_M = k_I\sigma_H\rho_{I_MH} = k_I\sigma_H\frac{\mathbf{a}'\mathbf{Z}_M\boldsymbol{\beta}}{\sqrt{\mathbf{a}'\mathbf{Z}_M\mathbf{a}}\sqrt{\boldsymbol{\beta}'\mathbf{T}_M\boldsymbol{\beta}}}, \tag{4.5}$$

where k_I is the standardized selection differential of the LMSI, $\sigma_H = \sqrt{\mathbf{a}'\mathbf{Z}_M\mathbf{a}}$ and $\sqrt{\boldsymbol{\beta}'\mathbf{T}_M\boldsymbol{\beta}}$ are the standard deviations of the variances of H and I_M, whereas ρ_{I_MH} and $\mathbf{a}'\mathbf{Z}_M\boldsymbol{\beta}$ are the correlation and the covariance between H and I_M respectively; $\mathbf{T}_M = Var\begin{bmatrix}\mathbf{y}\\ \mathbf{s}\end{bmatrix} = \begin{bmatrix}\mathbf{P} & \mathbf{S}\\ \mathbf{S} & \mathbf{S}\end{bmatrix}$ and $\mathbf{Z}_M = Var\begin{bmatrix}\mathbf{g}\\ \mathbf{s}\end{bmatrix} = \begin{bmatrix}\mathbf{C} & \mathbf{S}\\ \mathbf{S} & \mathbf{S}\end{bmatrix}$ are block matrices of covariance where $\mathbf{P} = Var(\mathbf{y})$, $\mathbf{S} = Var(\mathbf{s})$, and $\mathbf{C} = Var(\mathbf{g})$ are the covariance matrices of phenotypic values ($\mathbf{y}$), the marker score ($\mathbf{s}$), and the genetic value ($\mathbf{g}$) respectively in the population. Vectors $\mathbf{a}$ and $\boldsymbol{\beta}$ were defined in Eqs. (4.3) and (4.4) respectively.

The LMSI expected genetic gain per trait can be written as

$$\mathbf{E}_M = k_I \frac{\mathbf{Z}_M \boldsymbol{\beta}}{\sqrt{\boldsymbol{\beta}' \mathbf{T}_M \boldsymbol{\beta}}}. \tag{4.6}$$

All the parameters in Eq. (4.6) were previously defined.

4.1.3 The Maximized LMSI Parameters

Suppose that $\mathbf{P}$, $\mathbf{S}$ and $\mathbf{C}$ are known matrices; then, matrices $\mathbf{T}_M$ and $\mathbf{Z}_M$ are known and, according to the LPSI theory (Chap. 2 for details), the LMSI vector of coefficients ($\boldsymbol{\beta}_M$) that maximizes $\rho_{I_M H}$, R_M, and $\mathbf{E}_M$ can be written as

$$\boldsymbol{\beta} = \mathbf{T}_M^{-1} \mathbf{Z}_M \mathbf{a}, \tag{4.7}$$

whence the maximized selection response and the maximized correlation (or LMSI accuracy) between H and I_M can be written as

$$R_M = k_I \sqrt{\boldsymbol{\beta}' \mathbf{T}_M \boldsymbol{\beta}}, \tag{4.8a}$$

and

$$\rho_{I_M H} = \frac{\sigma_{I_M}}{\sigma_H}, \tag{4.8b}$$

respectively, where $\sigma_{I_M} = \sqrt{\boldsymbol{\beta}' \mathbf{T}_M \boldsymbol{\beta}}$ is the standard deviation of the variance of I_M and $\sigma_H = \sqrt{\mathbf{a}' \mathbf{Z}_M \mathbf{a}}$ is the deviation of the variance of H. Equations (4.8a) and (4.8b) show that the LMSI is a direct application of the LPSI theory in the marker-assisted selection (MAS) context.

Let $\mathbf{Q} = \mathbf{T}_M^{-1} \mathbf{Z}_M$; then, matrix $\mathbf{Q}$ can be written as

$$\mathbf{Q} = \begin{bmatrix} (\mathbf{P} - \mathbf{S})^{-1} (\mathbf{C} - \mathbf{S}) & \mathbf{0} \\ \mathbf{I} - (\mathbf{P} - \mathbf{S})^{-1} (\mathbf{C} - \mathbf{S}) & \mathbf{I} \end{bmatrix}, \tag{4.9}$$

whence $\boldsymbol{\beta} = \mathbf{Q}\mathbf{a}$, and as $\mathbf{w}_2' = [\, 0_1 \quad \cdots \quad 0_t \,]$, we can write the two vectors of $\boldsymbol{\beta}' = \left[\boldsymbol{\beta}_y' \quad \boldsymbol{\beta}_s' \right]$ as

$$\boldsymbol{\beta}_y = (\mathbf{P} - \mathbf{S})^{-1} (\mathbf{C} - \mathbf{S}) \mathbf{w} \quad \text{and} \quad \boldsymbol{\beta}_s = \left[\mathbf{I} - (\mathbf{P} - \mathbf{S})^{-1} (\mathbf{C} - \mathbf{S}) \right] \mathbf{w}. \tag{4.10a}$$

Another way of writing the marker score vector weights is

$$\boldsymbol{\beta}_s = \mathbf{w} - \boldsymbol{\beta}_y, \tag{4.10b}$$

where $\boldsymbol{\beta}_y = (\mathbf{P} - \mathbf{S})^{-1} (\mathbf{C} - \mathbf{S}) \mathbf{w}$. By Eq. (4.10b), the optimal LMSI can be written as

$$I_M = \mathbf{w}'\mathbf{s} + \boldsymbol{\beta}_y'(\mathbf{y} - \mathbf{s}). \tag{4.11}$$

Equation (4.11) indicates that, in practice, to estimate the optimal LMSI, we only need to estimate the vector of coefficients $\boldsymbol{\beta}_y$. By Eq. (4.10a), Eq. (4.8a) can be written as

$$R_M = k_I \sqrt{\mathbf{w}'\mathbf{C}(\mathbf{P} - \mathbf{S})^{-1}(\mathbf{C} - \mathbf{S})\mathbf{w} + \mathbf{w}'\mathbf{S}\left[\mathbf{I} - (\mathbf{P} - \mathbf{S})^{-1}(\mathbf{C} - \mathbf{S})\right]\mathbf{w}}. \tag{4.12}$$

Thus, by Eqs. (4.10a) and (4.12), when $\mathbf{S}$ is a null matrix, vector $\boldsymbol{\beta}_y$ is equal to $\boldsymbol{\beta}_y = \mathbf{P}^{-1}\mathbf{C}\mathbf{w} = \mathbf{b}$ and $R_M = k_I\sqrt{\mathbf{b}'\mathbf{P}\mathbf{b}} = R_I$, which are the LPSI vector of coefficients and its selection response respectively.

Assume that when the number of markers and genotypes tend to infinity, $\mathbf{S}$ tends to $\mathbf{C}$; then, at the limit, we can suppose that $\mathbf{S} = \mathbf{C}$, and by this latter result, R_M is equal to

$$k_I\sqrt{\mathbf{w}'\mathbf{C}\mathbf{w}}. \tag{4.13}$$

That is, Eq. (4.13) is the maximum value of the LMSI selection response when the numbers of markers and genotypes tend to infinity. Thus, the possible LMSI selection response values of Eq. (4.12) should be between $k_I\sqrt{\mathbf{b}'\mathbf{P}\mathbf{b}}$ and $k_I\sqrt{\mathbf{w}'\mathbf{C}\mathbf{w}}$, i.e.,

$$k_I\sqrt{\mathbf{b}'\mathbf{P}\mathbf{b}} \leq R_M \leq k_I\sqrt{\mathbf{w}'\mathbf{C}\mathbf{w}}, \tag{4.14}$$

or between 1 and $\frac{\sqrt{\mathbf{w}'\mathbf{C}\mathbf{w}}}{\sqrt{\mathbf{b}'\mathbf{P}\mathbf{b}}} = \frac{\sigma_H}{\sigma_I}$, that is,

$$1 \leq R_M \leq \frac{\sigma_H}{\sigma_I}. \tag{4.15}$$

Note that $\frac{\sigma_H}{\sigma_I} = \frac{1}{\rho_{HI}}$, where ρ_{HI} is the maximized correlation between the net genetic merit (H) and the LPSI (I) described in Chap. 2. Equation (4.15) indicates that LMSI efficiency tends to infinity when the ρ_{HI} value tends to zero and is an additional way of denoting the paradox of LMSI efficiency described by Knapp (1998), which implies that LMSI efficiency tends to infinity when the ρ_{HI} value tends to zero.

4.1.4 The LMSI for One Trait

For the one-trait case, matrices $\mathbf{T}_M$, $\mathbf{Z}_M$, and $\mathbf{Q}$ can be written as

$$\mathbf{T}_M = \begin{bmatrix} \sigma_y^2 & \sigma_s^2 \\ \sigma_s^2 & \sigma_s^2 \end{bmatrix}, \quad \mathbf{Z}_M = \begin{bmatrix} \sigma_g^2 & \sigma_s^2 \\ \sigma_s^2 & \sigma_s^2 \end{bmatrix} \quad \text{and} \quad \mathbf{Q} = \begin{bmatrix} \dfrac{\sigma_g^2 - \sigma_s^2}{\sigma_y^2 - \sigma_s^2} & 0 \\ \dfrac{\sigma_y^2 - \sigma_g^2}{\sigma_y^2 - \sigma_s^2} & 1 \end{bmatrix}, \tag{4.16}$$

where σ_y^2, σ_g^2, and σ_s^2 are the phenotypic, genetic, and marker score variances respectively. By Eqs. (4.10a) and (4.10b), when $\mathbf{a}' = [\,1 \quad 0\,]$, the elements of vector $\boldsymbol{\beta} = \mathbf{Qa}$ are

$$\beta_y = \frac{\sigma_g^2 - \sigma_s^2}{\sigma_y^2 - \sigma_s^2} \quad \text{and} \quad \beta_s = 1 - \beta_y, \tag{4.17a}$$

whence the optimal LMSI can be written as

$$I_M = s + \beta_y(y - s); \tag{4.17b}$$

whereas by Eq. (4.12), the maximized LMSI selection response can be written as

$$R_M = k_I \sqrt{\frac{\sigma_g^2\left(\sigma_g^2 - \sigma_s^2\right) + \sigma_s^2\left(\sigma_y^2 - \sigma_g^2\right)}{\sigma_y^2 - \sigma_s^2}}. \tag{4.18}$$

When $\sigma_s^2 = 0, \beta_y = \dfrac{\sigma_g^2}{\sigma_y^2} = h^2, I_M = h^2 y$, and $R_M = k\dfrac{\sigma_g^2}{\sigma_y} = k\sigma_y h^2 = R$, the selection response for the one-trait case without markers.

4.1.5 *Efficiency of LMSI Versus LPSI Efficiency for One Trait*

Suppose that the intensity of selection is the same in both indices; then, to compare LMSI versus LPSI efficiency for predicting the net genetic merit, we can use the ratio $\lambda_M = \frac{\rho_{I_M H}}{\rho_{HI}} = \frac{R_M}{R_I}$ (Bulmer 1980; Moreau et al. 1998), where R_I is the maximized LPSI selection response. In percentage terms, the LMSI versus LPSI efficiency can be written as

$$p_M = 100(\lambda_M - 1). \tag{4.19}$$

When $p_M = 0$, the efficiency of both indices is the same; when $p_M > 0$, the efficiency of the LMSI is higher than that of the LPSI, and when $p_M < 0$, LPSI efficiency is higher than LMSI efficiency for predicting the net genetic merit.

In the case of one trait, Lande and Thompson (1990) showed that LMSI efficiency (not in percentage terms) with respect to phenotypic efficiency can be written as

$$\lambda_M = \frac{R_M}{R} = \sqrt{\frac{q}{h^2} + \frac{(1-q)^2}{1-qh^2}}, \tag{4.20}$$

where R_M was defined in Eq. (4.18), $R = k\sigma_y h^2$, h^2 is the trait heritability, and $q = \frac{\sigma_s^2}{\sigma_g^2}$ is the proportion of additive genetic variance explained by the markers. According to Eq. (4.20), the advantage of the LMSI over phenotypic selection increases as the population size increases and heritability decreases, because in such cases, $q = \frac{\sigma_s^2}{\sigma_g^2}$ tends to 1 and Eq. (4.20) approaches $\frac{1}{h}$. Therefore, the LMSI is most efficient for traits with low heritability and when the marker score explains a large proportion of the genetic variance. Thus, note that when h^2 tends to zero, $\frac{1}{h}$ tends to infinity; this means that in the asymptotic context, LMSI efficiency with respect to phenotypic efficiency for one trait (Eq. 4.20) tends to infinity and this is the LMSI paradox pointed out by Knapp (1998). There are other problems associated with the LMSI: it increases the selection response only in the short term and can result in lower cumulative responses in the longer term than phenotypic selection, as the LMSI fixes the QTL at a faster rate than phenotypic selection. In addition, it requires the weights (Eq. 4.17a) to be updated, because in each generation the frequency of the QTL changes (Dekkers and Settar 2004).

4.1.6 Statistical LMSI Properties

Assume that H and I_M have bivariate joint normal distribution, $\boldsymbol{\beta} = \mathbf{T}_M^{-1}\mathbf{Z}_M\mathbf{a}$, and that $\mathbf{P}$, $\mathbf{C}$, $\mathbf{S}$, and $\mathbf{w}$ are known; then, the statistical LMSI properties are the same as the LPSI properties described in Chap. 2. That is,

1. $\sigma_{I_M}^2 = \sigma_{HI_M}$: the variance of I_M ($\sigma_{I_M}^2$) and the covariance between H and I_M (σ_{HI_M}) are the same.
2. The maximized correlation between H and I_M (or I_M accuracy) is $\rho_{HI_M} = \frac{\sigma_{I_M}}{\sigma_H}$.
3. The variance of the predicted error, $Var(H - I_M) = \left(1 - \rho_{HI_M}^2\right)\sigma_H^2$, is minimal.
4. The total variance of H explained by I_M is $\sigma_{I_M}^2 = \rho_{HI_M}^2\sigma_H^2$.
5. The heritability of I_M is $h_M^2 = \frac{\boldsymbol{\beta}_M'\mathbf{Z}_M\boldsymbol{\beta}_M}{\boldsymbol{\beta}_M'\mathbf{T}_M\boldsymbol{\beta}_M}$.

Properties 1 to 4 are the same as LPSI properties 1 to 4, but, because the LMSI jointly incorporates the phenotypic and marker information to predict the net genetic merit, LMSI accuracy should be higher than LPSI accuracy. The same is true of the LMSI selection response and expected genetic gain per trait when compared with the LPSI selection response and expected genetic gain per trait.

4.2 The Genome-Wide Linear Selection Index

The genome-wide linear marker selection index (GW-LMSI) is a single-stage procedure that treats information at each individual marker as a separate trait. Thus, all marker information can be entered together with phenotypic information into the GW-LMSI, which is then used to predict the net genetic merit. In a similar manner to the LMSI, the GW-LMSI exploits the linkage disequilibrium between markers and the QTL produced when inbred lines are crossed.

4.2.1 The GW-LMSI Parameters

In a similar manner to the LPSI, the main objective of the GW-LMSI is to predict the net genetic merit values of each individual and select the best individuals for further breeding. In the GW-LMSI context, the net genetic merit can be written as

$$H = \mathbf{w}'\mathbf{g} + \mathbf{w}_2'\mathbf{m} = \begin{bmatrix} \mathbf{w}' & \mathbf{w}_2' \end{bmatrix} \begin{bmatrix} \mathbf{g} \\ \mathbf{m} \end{bmatrix} = \mathbf{a}_W'\mathbf{z}_W, \tag{4.21}$$

where $\mathbf{g}' = [\, g_1 \quad \ldots \quad g_t \,]$ ($j = 1, 2, \ldots, t$ = number of traits) is the vector of breeding values, $\mathbf{w}' = [\, w_1 \quad \cdots \quad w_t \,]$ is the vector of economic weights associated with the breeding values, and $\mathbf{w}_2' = [\, 0_1 \quad \cdots \quad 0_m \,]$ is a null vector associated with the coded values of the markers $\mathbf{m}' = [\, m_1 \quad \cdots \quad m_m \,]$, where m_j ($j = 1, 2, \ldots, m$ = number of markers) is the jth marker in the training population; $\mathbf{a}_W' = [\, \mathbf{w}' \quad \mathbf{w}_2' \,]$ and $\mathbf{z}_W = [\, \mathbf{g}' \quad \mathbf{m}' \,]$.

The GW-LMSI (I_W) combines the phenotypic value and the molecular information linked to the individual traits to predict H values in each selection cycle. It can be written as

$$I_W = \boldsymbol{\beta}_y'\mathbf{y} + \boldsymbol{\beta}_m'\mathbf{m} = \begin{bmatrix} \boldsymbol{\beta}_y' & \boldsymbol{\beta}_m' \end{bmatrix} \begin{bmatrix} \mathbf{y} \\ \mathbf{m} \end{bmatrix} = \boldsymbol{\beta}_W'\mathbf{t}_W, \tag{4.22}$$

where $\boldsymbol{\beta}_y'$ and $\boldsymbol{\beta}_m$ are vectors of phenotypic and marker weights respectively; $\mathbf{y}' = [\, y_1 \quad \cdots \quad y_t \,]$ is the vector of phenotypic values and $\mathbf{m}$ was defined in Eq. (4.21); $\boldsymbol{\beta}_W' = \begin{bmatrix} \boldsymbol{\beta}_y' & \boldsymbol{\beta}_m' \end{bmatrix}$ and $\mathbf{t}_W' = [\, \mathbf{y}' \quad \mathbf{m}' \,]$.

The GW-LSI selection response can be written as

$$R_W = k_I \sigma_H \rho_{I_W H} = k_I \sigma_H \frac{\mathbf{a}_W' \boldsymbol{\Psi} \boldsymbol{\beta}_W}{\sqrt{\mathbf{a}_W' \boldsymbol{\Psi} \mathbf{a}_W} \sqrt{\boldsymbol{\beta}_W' \boldsymbol{\Phi} \boldsymbol{\beta}_W}}, \tag{4.23a}$$

where k_I is the standardized selection differential of the GW-LMSI, $\sigma_H^2 = \mathbf{a}_W' \boldsymbol{\Psi} \mathbf{a}_W$ and $Var(I_W) = \boldsymbol{\beta}_W' \boldsymbol{\Phi} \boldsymbol{\beta}_W$ are the variance of H and I_W, whereas $\rho_{I_W H} = \frac{\mathbf{a}_W' \boldsymbol{\Psi} \boldsymbol{\beta}_W}{\sqrt{\mathbf{a}_W' \boldsymbol{\Psi} \mathbf{a}_W} \sqrt{\boldsymbol{\beta}_W' \boldsymbol{\Phi} \boldsymbol{\beta}_W}}$ and $\mathbf{a}_W' \boldsymbol{\Psi} \boldsymbol{\beta}_W$ are the correlation and the covariance between

H and I_W respectively; $\mathbf{\Phi} = Var\begin{bmatrix}\mathbf{y}\\ \mathbf{m}\end{bmatrix} = \begin{bmatrix}\mathbf{P} & \mathbf{W}'\\ \mathbf{W} & \mathbf{M}\end{bmatrix}$ and $\mathbf{\Psi} = Var\begin{bmatrix}\mathbf{g}\\ \mathbf{m}\end{bmatrix} = \begin{bmatrix}\mathbf{C} & \mathbf{W}'\\ \mathbf{W} & \mathbf{M}\end{bmatrix}$ are block covariance matrices where $\mathbf{P} = Var(\mathbf{y})$, $\mathbf{M} = Var(\mathbf{m})$, $\mathbf{C} = Var(\mathbf{g})$, and $\mathbf{W} = Cov(\mathbf{y}, \mathbf{m}) = Cov(\mathbf{g}, \mathbf{m})$ are the covariance matrices of phenotypic values ($\mathbf{y}$), the molecular marker ($\mathbf{m}$) coded values, and the genetic ($\mathbf{g}$) values, whereas $\mathbf{W}$ is the covariance matrix between $\mathbf{y}$ and $\mathbf{m}$, and between $\mathbf{g}$ and $\mathbf{m}$. The size of matrices $\mathbf{P}$ and $\mathbf{C}$ is $t \times t$, but the sizes of matrices $\mathbf{M}$ and $\mathbf{W}$ are $m \times m$ and $m \times t$ respectively.

From a theoretical point of view, Crossa and Cerón-Rojas (2011) showed that matrix $\mathbf{M}$ can be written as

$$\mathbf{M} = \begin{bmatrix} 1 & (1-2\delta_{11}) & \cdots & (1-2\delta_{1N}) \\ (1-2\delta_{21}) & 1 & \cdots & (1-2\delta_{2N}) \\ \vdots & \vdots & \ddots & \vdots \\ (1-2\delta_{N1}) & (1-2\delta_{N2}) & \cdots & 1 \end{bmatrix}, \tag{4.23b}$$

where $(1 - 2\delta_{ij})$ is the covariance (or correlation) and δ_{ij} the recombination frequency between the ith and jth marker ($i, j = 1, 2, \ldots, m$ = number of markers). According to Crossa and Cerón-Rojas (2011), matrix $\mathbf{W}$ can be written as

$$\mathbf{W} = \begin{bmatrix} (1-2r_{11})\alpha_{11} & (1-2r_{11})\alpha_{12} & \cdots & (1-2r_{1N})\alpha_{1N_Q} \\ (1-2r_{21})\alpha_{21} & (1-2r_{22})\alpha_{22} & \cdots & (1-2r_{2N})\alpha_{2N_Q} \\ \vdots & \vdots & \ddots & \vdots \\ (1-2r_{t1})\alpha_{t1} & (1-2r_{N2})\alpha_{t2} & \cdots & (1-2r_{NN})\alpha_{tN_Q} \end{bmatrix}, \tag{4.23c}$$

where $(1 - 2r_{ik})\alpha_{qk}$ ($i = 1, 2, \ldots, m$, $k = 1, 2, \ldots, N_Q$ = number of QTL, $q = 1, 2, \ldots, t$) is the covariance between the qth trait and the ith marker; r_{ik} is the recombination frequency between the ith marker and the kth QTL; and α_{qk} is the effect of the kth QTL over the qth trait.

The GW-LMSI expected genetic gain per trait can be written as

$$\mathbf{E}_{LW} = k_I \frac{\mathbf{\Psi\beta}}{\sqrt{\mathbf{\beta}'\mathbf{\Phi\beta}}}. \tag{4.24}$$

All parameters in Eq. (4.24) were previously defined.

Matrix $\mathbf{\Phi}$ could be singular, i.e., its inverse ($\mathbf{\Phi}^{-1}$) could not exist because matrix $\mathbf{W}$ is singular. Suppose that matrices $\mathbf{\Phi}$ and $\mathbf{\Psi}$ are known; then, according to the LPSI theory, the GW-LMSI vector of coefficients ($\boldsymbol{\beta}_W$) that maximizes $\rho_{I_W H}$ can be written as

$$\boldsymbol{\beta}_W = \boldsymbol{\Phi}^{-}\boldsymbol{\Psi}\mathbf{a}_W, \tag{4.25a}$$

where matrix $\boldsymbol{\Phi}^{-}$ denotes a generalized inverse of $\boldsymbol{\Phi}$. By Eq. (4.25a), the maximized GW-LMSI selection response is

$$R_W = k_I\sqrt{\boldsymbol{\beta}'_W\boldsymbol{\Phi}\boldsymbol{\beta}_W}. \tag{4.25b}$$

Equations (4.25a) and (4.25b) show that the GW-LMSI is a direct application of the LPSI to MAS. By Eq. (4.25a), the maximized correlation between H and I_W is

$$\rho_{I_WH} = \frac{\sigma_{I_W}}{\sigma_H}, \tag{4.25c}$$

where $\sigma_{I_W} = \sqrt{\boldsymbol{\beta}'_W\boldsymbol{\Phi}\boldsymbol{\beta}_W}$ is the standard deviation of the variance of I_W and $\sigma_H = \sqrt{\mathbf{a}'_W\boldsymbol{\Psi}\mathbf{a}_W}$ is the standard deviation of the variance of H.

4.2.2 Relationship Between the GW-LMSI and the LPSI

Matrix $\boldsymbol{\Phi}^{-}$ can be written as

$$\boldsymbol{\Phi}^{-} = \begin{bmatrix} \mathbf{L}^{-} & -\mathbf{L}^{-}\mathbf{W}'\mathbf{M}^{-} \\ -\mathbf{M}^{-}\mathbf{W}\mathbf{L}^{-} & \mathbf{M}^{-} + \mathbf{M}^{-}\mathbf{W}\mathbf{L}^{-}\mathbf{W}'\mathbf{M}^{-} \end{bmatrix}, \tag{4.26}$$

where $\mathbf{L}^{-}$ is a generalized inverse of matrix $\mathbf{L} = \mathbf{P} - \mathbf{W}'\mathbf{M}^{-}\mathbf{W}$, and $\mathbf{M}^{-}$ is a generalized inverse of matrix $\mathbf{M}$. In matrix $\boldsymbol{\Phi}^{-}$, the inverse of matrix $\mathbf{W}$ is not required and the standard inverse of matrix $\mathbf{M}$ ($\mathbf{M}^{-1}$) may exist. In the latter case, the standard inverse of matrix $\mathbf{L}$ ($\mathbf{L}^{-1}$) exists and can be written as $\mathbf{L}^{-1} = (\mathbf{P} - \mathbf{W}'\mathbf{M}^{-1}\mathbf{W})^{-1} = \mathbf{P}^{-1} + \mathbf{P}^{-1}\mathbf{W}'[\mathbf{M} - \mathbf{W}\mathbf{P}^{-1}\mathbf{W}']^{-1}\mathbf{W}\mathbf{P}^{-1}$ (Searle et al. 2006).

By Eq. (4.26) and because $\mathbf{w}'_2 = [0_1 \quad \cdots \quad 0_N]$, the vector components of $\boldsymbol{\beta}'_W = [\boldsymbol{\beta}'_y \quad \boldsymbol{\beta}'_m]$, or $\boldsymbol{\beta}_W = \boldsymbol{\Phi}^{-}\boldsymbol{\Psi}\mathbf{a}_W$, can be written as

$$\boldsymbol{\beta}_y = [\mathbf{L}^{-}\mathbf{C} - \mathbf{L}^{-}\mathbf{W}'\mathbf{M}^{-}\mathbf{W}]\mathbf{w} \tag{4.27}$$

and

$$\boldsymbol{\beta}_m = [(\mathbf{M}^{-} + \mathbf{M}^{-}\mathbf{W}\mathbf{L}^{-}\mathbf{W}'\mathbf{M}^{-})\mathbf{W} - \mathbf{M}^{-}\mathbf{W}\mathbf{L}^{-}\mathbf{C}]\mathbf{w}, \tag{4.28}$$

where $\mathbf{w}$ is the vector of economic weights. Suppose that there is no marker information; then, matrices $\mathbf{M}$ and $\mathbf{W}$ are null and Eq. (4.27) is equal to $\boldsymbol{\beta}_y = \mathbf{P}^{-1}\mathbf{C}\mathbf{w} = \mathbf{b}$ (the LPSI vector of coefficients), whereas $\boldsymbol{\beta}_m = \mathbf{0}$ and $R_W = k_I\sqrt{\boldsymbol{\beta}'_W\boldsymbol{\Phi}\boldsymbol{\beta}_W} = k_I\sqrt{\mathbf{b}'\mathbf{P}\mathbf{b}} = R_I$, the LPSI selection response. Now suppose that the markers explain all the genetic variability; in this case, $\boldsymbol{\beta}_y = \mathbf{0}$ and $\boldsymbol{\beta}_m = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Y}$, the matrix of linear regression coefficients in the multivariate context,

where $(\mathbf{X}'\mathbf{X})^{-}$ is a generalized inverse matrix of $\mathbf{X}'\mathbf{X}$ and $\mathbf{Y}$ is a matrix of phenotypic observations.

4.2.3 Statistical Properties of GW-LMSI

Assume that H and I_W have bivariate joint normal distribution, $\boldsymbol{\beta}_W = \boldsymbol{\Phi}^{-}\boldsymbol{\Psi}\mathbf{a}_W$, and $\mathbf{P}$, $\mathbf{C}$, $\mathbf{M}$, $\mathbf{W}$, and $\mathbf{w}$ are known; then, the statistical GW-LMSI properties are the same as the LMSI properties. That is,

1. $\sigma^2_{I_W} = \sigma_{HI_W}$, i.e., the variance of I_W ($\sigma^2_{I_W}$) and the covariance between H and I_W (σ_{HI_W}) are the same.
2. The maximized correlation between H and I_W, or I_W accuracy, is $\rho_{HI_W} = \frac{\sigma_{I_W}}{\sigma_H}$.
3. The variance of the predicted error, $Var(H - I_W) = \left(1 - \rho^2_{HI_W}\right)\sigma^2_H$, is minimal.
4. The total variance of H explained by I_W is $\sigma^2_{I_W} = \rho^2_{HI_W}\sigma^2_H$.

According to Lange and Whittaker (2001), GW-LMSI efficiency should be greater than LMSI efficiency. However, this would be true only if matrices $\mathbf{P}$, $\mathbf{C}$, $\mathbf{M}$, and $\mathbf{W}$ are known and trait heritability is very low.

4.3 Estimating the LMSI Parameters

When covariance matrices $\mathbf{P}$, $\mathbf{C}$, and $\mathbf{S}$, and the vector of economic weights ($\mathbf{w}$) are known, there is no error in the estimation of the LMSI parameters (selection response, expected genetic gain, etc.); the same is true for the GW-LMSI when, in addition to $\mathbf{P}$, $\mathbf{C}$, and $\mathbf{w}$, the covariance matrices $\mathbf{M}$ and $\mathbf{W}$ are known. In such cases, the relative efficiency of the LMSI (GW-LMSI) depends only on the heritability of the traits and on the portion of phenotypic variation associated with markers. Using simulated data, Lange and Whittaker (2001) found that GW-LMSI efficiency was higher than LMSI efficiency when trait heritability was 0.2 and matrices $\mathbf{P}$, $\mathbf{C}$, $\mathbf{M}$, and $\mathbf{W}$ were known. When $\mathbf{P}$, $\mathbf{C}$, $\mathbf{S}$, $\mathbf{M}$, and $\mathbf{W}$ are unknown, it is necessary to estimate them; then, the LMSI and GW-LMSI vector of coefficients and the effects associated with markers are estimated with some error. This error leads to lower LMSI and GW-LMSI efficiency than expected under the assumption that the parameters are known; however, in the latter case, Lange and Whittaker (2001) also found that GW-LMSI efficiency was greater than that of the LMSI when trait heritability was 0.05. Moreover, in the LMSI there is additional bias in the estimation of the parameters because only markers with significant effects are included in the index (Moreau et al. 1998).

In Chap. 2, we described the restricted maximum likelihood (REML) method for estimating matrices $\mathbf{P}$ and $\mathbf{C}$. Some authors (Lande and Thompson 1990; Charcosset

and Gallais 1996; Hospital et al. 1997; Moreau et al. 1998, 2007) have described methods for estimating marker scores, the variance of the marker scores, the LMSI vector of coefficients, etc., in the context of one trait; however, up to now there have been no reports on the estimation of matrix **S** in the multi-trait case. Lange and Whittaker (2001) only indicated that matrix **S** can be estimated as $\widehat{\mathbf{S}} = Var(\widehat{\mathbf{s}})$, where $\widehat{\mathbf{s}}$ is a vector of estimated marker scores associated with several individual traits.

The main problems associated with the estimated LMSI parameters are:

1. The estimated values of the covariance matrix **S** ($\widehat{\mathbf{S}}$) tend to overestimate the genetic covariance matrix (**C**).
2. The estimated variances of the marker scores can be negative.

When the first point is true, the estimated LMSI selection response and efficiency could be negative because the estimated matrix $\widehat{\mathbf{T}}_M = \begin{bmatrix} \widehat{\mathbf{P}} & \widehat{\mathbf{S}} \\ \widehat{\mathbf{S}} & \widehat{\mathbf{S}} \end{bmatrix}$ is not positive definite (all eigenvalues positive) and the estimated matrix $\widehat{\mathbf{Z}}_M = \begin{bmatrix} \widehat{\mathbf{G}} & \widehat{\mathbf{S}} \\ \widehat{\mathbf{S}} & \widehat{\mathbf{S}} \end{bmatrix}$ is not positive semi-definite (no negative eigenvalues). In addition, the results can lead to all weights being placed on the molecular score and the weights on the phenotype values can be negative (Moreau et al. 2007). When the second point is true, the variance of the marker scores is not useful. The two problems indicated above could be caused by using the same data set to select markers and to estimate marker effects, and there is no simple way of solving them. Lande and Thompson (1990) proposed that the markers used to obtain $\widehat{\mathbf{S}}$ be selected a priori as those with the most highly significant partial regression coefficients from among all the markers in the linkage group analyzed in the previous generation. Zhang and Smith (1992, 1993) proposed using two independent sets of markers: one to estimate marker effects and the other to select markers. Additional solutions to these problems were described by Moreau et al. (2007).

In this subsection, we describe methods (in the univariate and multivariate context) for estimating molecular marker effects, marker scores, and their variance and covariance, and for estimating the LMSI and GW-LMSI vector of coefficients, selection response, expected genetic gain, and accuracy. This subsection is only for illustration; we use the same data set to select markers, and to estimate marker effects and the variance of marker scores.

4.3.1 Estimating the Marker Score

According to Eqs. (4.11) and (4.17b), when the vector of economic weights is equal to $\mathbf{a}' = [\,1 \quad 0\,]$, the LMSI for the ith trait y_i ($i = 1, 2, \cdots, t$; t = number of traits) value can be written as $I_{M_{li}} = s_i + \beta_{y_i}(y_i - s_i)$ ($l = 1, 2, \cdots, n$; n = number of

individuals or genotypes), where $\beta_{yi} = \dfrac{\sigma^2_{g_i} - \sigma^2_{s_i}}{\sigma^2_{y_i} - \sigma^2_{s_i}} = \dfrac{h_i^2(1-q_i)}{1-q_i h_i^2}$ is the LMSI coefficient, $h_i^2 = \dfrac{\sigma^2_{g_i}}{\sigma^2_{y_i}}$ is the heritability of the ith trait, and $q_i = \dfrac{\sigma^2_{s_i}}{\sigma^2_{g_i}}$ is the proportion of genetic variance explained by the QTL or markers associated with the ith trait; $s_i = \sum_{j=1}^{M} \theta_j x_j$ ($j = 1, 2, \cdots, M$; M = number of selected markers) is the ith individual trait marker score; and $\sigma^2_{y_i}$, $\sigma^2_{g_i}$, and $\sigma^2_{s_i}$ are the ith variances of the phenotypic, genetic, and marker score values respectively.

The simplest way of estimating the ith marker score s_i is to perform a multiple linear regression of phenotypic values (y_i) on the coded values of the markers (x_j) and then select the markers statistically linked to the ith QTL that explain most of the variability in the regression model and use them to construct $s_i = \sum_{j \in M} \theta_j x_j$.

We can fit the model $y_i^* = \sum_{j \in M} \theta_j x_j + e$, where $y_i^* = y_i - \bar{y}_i$ and $\bar{y}_i$ are the average values of the ith trait, by maximum likelihood or least squares. When estimating θ_j, the main problem is to choose the set of markers M based on criteria for declaring markers as significant and then use the estimated values of θ_j ($\widehat{\theta}_j$) to estimate the ith marker score s_i as $\widehat{s}_i = \sum_{j \in M} \widehat{\theta}_j x_j$. The values of $\widehat{s}_i$ may increase or decrease according to the number of markers (x_j) included in the model, and $\widehat{s}_i$ affects LMSI selection response and efficiency by means of the estimated variance of $\widehat{s}_i$ ($\widehat{\sigma}^2_{\widehat{s}_i}$) (Figs. 4.1 and 4.2).

According to the least squares method of estimation, $\widehat{\boldsymbol{\theta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}^*$ is an estimator of the vector of regression coefficients $\boldsymbol{\theta}' = [\theta_1 \quad \theta_2 \quad \cdots \quad \theta_m]$, where m ($m < n$) is the number of markers, $\mathbf{X}$ is a matrix $n \times m$ of coded marker values (e.g., 1, 0 and -1 for marker genotypes *AA*, *Aa*, and *aa* respectively) and $\mathbf{y}^*$ is a vector $n \times 1$ of phenotypic values centered based on its average values. Only a subset $M(M < m)$ of the m markers is statistically linked to the QTL and then only a subset M of the estimated vector $\widehat{\boldsymbol{\theta}}$ values is selected to estimate s_i as $\widehat{s}_i = \sum_{j=1}^{M} \widehat{\theta}_j x_j$.

To illustrate how to obtain $\widehat{s}_i = \sum_{j \in M} \widehat{\theta}_j x_j$, we use a real maize (*Zea mays*) F_2 population with 247 genotypes (each one with two repetitions), 195 molecular markers, and four traits – grain yield (GY, ton ha^{-1}); plant height (PHT, cm), ear height (EHT, cm), and anthesis day (AD, days) – evaluated in one environment. In an F_2 population, the marker homozygous loci for the allele from the first parental line can be coded by 1, whereas the marker homozygous loci for the allele from the second parental line can be coded by -1, and the marker heterozygous loci by 0.

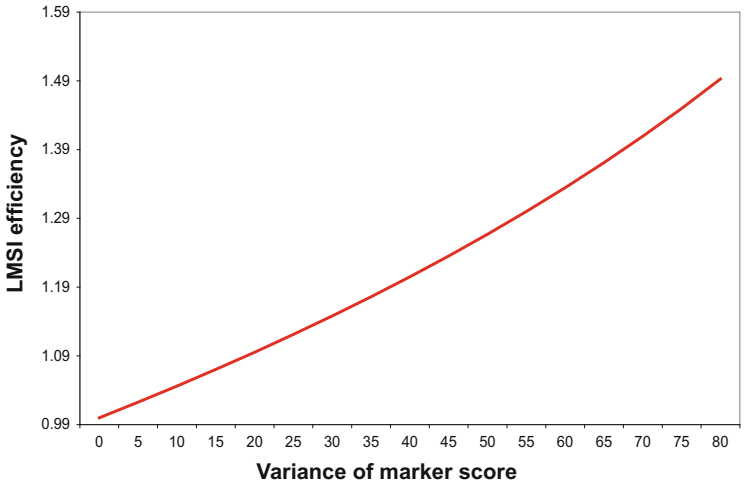

Fig. 4.1 Efficiency of the linear molecular selection index with respect to phenotypic selection for the one-trait case for different values of the variance of the marker score when the phenotypic and genetic variances are fixed

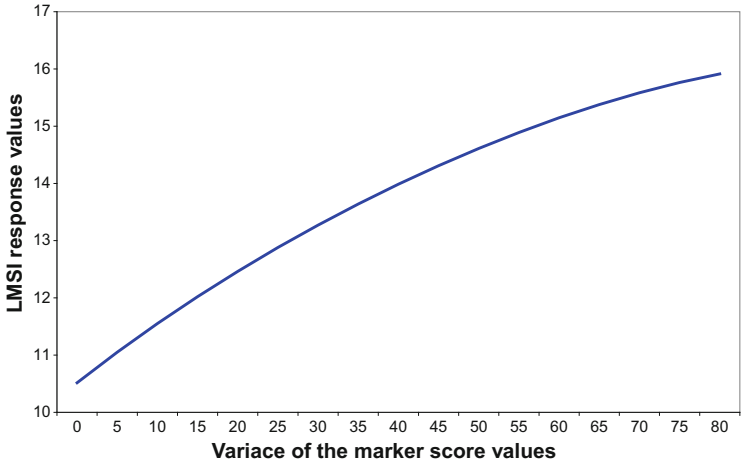

Fig. 4.2 Selection response values of the linear molecular selection index for the one-trait case for different values of the variance of the marker score when the phenotypic and genetic variances are fixed

For this example, we used trait PHT. Only seven markers were statistically linked to the PHT. The estimated vector of regression coefficients for these seven markers was $\widehat{\mathbf{b}}' = [5.46 \quad -4.54 \quad 0.98 \quad 7.39 \quad -7.75 \quad -1.91 \quad -3.53]$. Table 4.1 presents the first 20 genotypes, the coded values of the seven selected markers, and the first 20 estimated $\widehat{s}_{PHT}$ values of the 247 genotypes in the maize (*Zea mays*) F_2

Table 4.1 Number of selected genotypes, coded values of seven selected markers, and estimated marker score values obtained from a maize (*Zea mays*) F_2 population with 247 genotypes and 195 molecular markers

	Coded values of the selected markers							
Number of genotypes	M1	M2	M3	M4	M5	M6	M7	Marker score
1	0	0	0	0	0	1	−1	1.62
2	−1	−1	0	0	0	−1	0	0.99
3	0	0	0	0	0	0	1	−3.53
4	1	1	0	0	0	−1	−1	6.37
5	1	1	0	−1	−1	−1	−1	6.72
6	0	0	1	0	0	0	0	0.98
7	1	1	0	1	1	0	0	0.57
8	0	0	0	0	0	0	0	0
9	0	0	1	0	0	1	0	−0.93
10	0	0	1	1	0	0	1	4.84
11	0	0	0	0	0	0	0	0
12	−1	−1	0	0	0	0	0	−0.92
13	0	0	0	0	0	0	0	0
14	1	1	0	−1	−1	0	−1	4.81
15	0	0	1	−1	−1	0	0	1.34
16	0	0	0	0	0	0	0	0
17	−1	−1	0	0	0	0	1	−4.46
18	−1	−1	0	0	0	0	1	−4.46
19	−1	−1	1	0	0	−1	1	−1.56
20	0	0	0	0	0	0	−1	3.53

population. According to $\widehat{\boldsymbol{\theta}}'$ and the coded values of the seven markers, the first estimated $\widehat{s}_{PHT}$ value was obtained as $\widehat{s}_{PHT1} = -1.91(1) + -3.53(-1) = 1.62$; the second estimated $\widehat{s}_{PHT}$ value was obtained as $\widehat{s}_{PHT2} = 5.46(-1) + -4.54(-1) - 1.91(-1) = 0.99$, etc. The 20th estimated $\widehat{s}_{PHT}$ value was obtained as $\widehat{s}_{PHT20} = -3.53(-1) = 3.53$. This estimation procedure is valid for any number of genotypes and markers.

Figure 4.3 shows the distribution of the 247 estimated marker scores associated with traits PHT and EHT of the maize F_2 population. Note that the estimated marker score values approach normal distribution.

4.3.2 Estimating the Variance of the Marker Score

There are many methods of estimating the variance of the marker score associated with the *i*th trait ($\sigma^2_{s_i}$); the first one was proposed by Lande and Thompson (1990). According to these authors, $\sigma^2_{s_i}$ can be estimated as

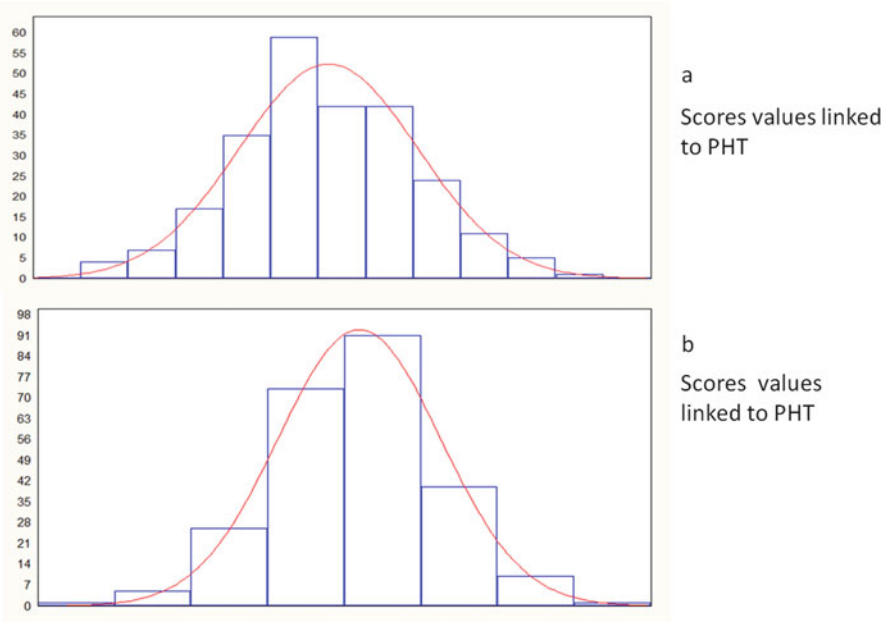

Fig. 4.3 Distribution of the marker scores associated with traits (**a**) plant height and (**b**) ear height of a maize (*Zea mays*) F_2 population. Note that the distribution of frequencies of the marker score values approaches normal distribution

$$\widehat{\sigma}^2_{\widehat{s}_i} = \widehat{\boldsymbol{\theta}}_i' \mathbf{M}_i \widehat{\boldsymbol{\theta}}_i - \frac{M\widehat{\sigma}^2_{e_i}}{n}, \tag{4.29}$$

where $\widehat{\boldsymbol{\theta}}_i$ is the estimated vector of regression coefficients of the selected markers, $\mathbf{M}_i = \frac{2}{n}\mathbf{X}_i'\mathbf{X}_i$ is the covariance matrix $M \times M$ of the selected markers that are statistically linked to the ith trait marker loci; $\widehat{\sigma}^2_{e_i} = \frac{\mathbf{y}'(\mathbf{I}-\mathbf{H})\mathbf{y}}{n-M-1}$ is the unbiased estimated variance of the residuals, $\mathbf{H} = \mathbf{I} - \mathbf{X}_i\left(\mathbf{X}_i'\mathbf{X}_i\right)^{-1}\mathbf{X}_i'$, $\mathbf{I}$ is an identity matrix $n \times n$, M is the number of selected markers statistically linked to the QTL, and $\mathbf{X}_i$ is a matrix $n \times M$ with the coded values of the selected markers. According to Lande and Thompson (1990), Eq. (4.29) is an unbiased estimator of $\sigma^2_{s_i}$ and its variance can be written as

$$Var\left(\widehat{\sigma}^2_{\widehat{s}_i}\right) = \frac{4\sigma^2_{s_i}\sigma^2_{e_i}}{n} + \frac{2M\left(\sigma^2_{e_i}\right)^2}{n^2} + \frac{2M^2\left(\sigma^2_{e_i}\right)^2}{n^2(n-M)}, \tag{4.30}$$

which tends to zero when n, the number of genotypes or individuals, is very high.

From Eq. (4.29), it is possible to obtain an estimator of the covariance between the *i*th and *j*th marker scores when the number of selected markers statistically linked to the QTL is the same in the *i*th and *j*th traits. Thus, by Eq. (4.29), the covariance between the *i*th and *j*th marker scores can be estimated as

$$\widehat{\sigma}_{\widehat{s}_{ij}} = \widehat{\boldsymbol{\theta}}_i' \mathbf{M}_{ij} \widehat{\boldsymbol{\theta}}_j - \frac{M \widehat{\sigma}_{e_{ij}}}{n}, \tag{4.31}$$

where $\widehat{\boldsymbol{\theta}}_i$ and $\widehat{\boldsymbol{\theta}}_j$ are the estimated vectors of regression coefficients of the selected markers associated with the *i*th and *j*th trait loci respectively; $\mathbf{M}_{ij} = \frac{2}{n}\mathbf{X}_i'\mathbf{X}_j$ is the covariance matrix $M \times M$ of the markers statistically linked to the *i*th and *j*th trait marker loci; $\mathbf{X}_i$ and $\mathbf{X}_j$ are $n \times M$ matrices with the coded values of the selected markers associated with the *i*th and *j*th trait loci respectively; $\widehat{\sigma}_{e_{ij}} = \dfrac{\mathbf{y}_i'(\mathbf{I} - \mathbf{H}_{ij})\mathbf{y}_j}{n - M - 1}$ is the estimated covariance of the residuals between the *i*th ($\mathbf{y}_i$) and *j*th ($\mathbf{y}_j$) trait values, $\mathbf{H}_{ij} = \mathbf{I} - \mathbf{X}_i\left(\mathbf{X}_i'\mathbf{X}_j\right)^{-1}\mathbf{X}_j'$, $\mathbf{I}$ is an identity matrix $n \times n$, and M is the number of selected markers statistically linked to the QTL.

According to the PHT values described in Sect. 4.3.1 of this chapter, $M = 7$, $n = 247$, $\widehat{\sigma}^2_{e_i} = 180.80$ and $\widehat{\sigma}^2_{\widehat{s}_{PHT}} = 48.23$ (Eq. 4.29). Note that $\widehat{\sigma}^2_{\widehat{s}_{PHT}} \leq \widehat{\sigma}^2_{g_{PHT}}$, where $\widehat{\sigma}^2_{g_{PHT}} = 83.0$ is an estimate of the genetic variance of PHT. The estimated portion of the genetic variance attributable to $\widehat{\sigma}^2_{\widehat{s}_{PHT}} = 48.23$ was $\widehat{q}_{PHT} = \dfrac{48.23}{83} = 0.5811$; that is, the seven markers explain 58.11% of the genetic variance associated with PHT.

Charcosset and Gallais (1996) considered two possible methods of estimating $\sigma^2_{s_i}$ based on the coefficient of multiple determination or squared multiple correlation R^2 (note that in this case R^2 is not the square of the selection response). The coefficient R^2 gives the portion of the total variation in the phenotypic values that is "explained" by, or attributable to, the markers and can be written as

$$R^2 = \frac{\widehat{\boldsymbol{\theta}}\mathbf{X}'\mathbf{y} - n\bar{y}^2}{\mathbf{y}'\mathbf{y} - n\bar{y}^2} = \frac{\widehat{\sigma}^2_s}{\widehat{\sigma}^2_y}, \tag{4.32a}$$

where $\widehat{\boldsymbol{\theta}}\mathbf{X}'\mathbf{y} - n\bar{y}^2$ is the overall regression sum of squares adjusted for the intercept and $\mathbf{y}'\mathbf{y} - n\bar{y}^2$ is the total sum of squares adjusted for the mean. The coefficient R^2 is equal to 1 if the fitted equation $y_i = \theta_0 + \sum_{j \in M} \theta_j x_j + e_i$ passes through all the data points, so that all residuals are null; then, the markers explain all the phenotypic variance. At the other extreme, R^2 is zero if $\bar{y}_i = \widehat{\theta}_0$ and the estimated regression coefficients are null, i.e., $\widehat{\theta}_1 = \widehat{\theta}_2 = \cdots = \widehat{\theta}_M = 0$. In the latter case, markers do not affect the phenotypic observations and the variance of the marker score values is zero. Thus, the R^2 values are between 0 and 1, i.e., $0 \leq R^2 \leq 1.0$. Equation (4.32a) is useful for estimating $\sigma^2_{s_i}$ as $\widehat{\sigma}^2_{y_i} \sum_{j=1}^{M} R_j^2 = \widehat{\sigma}^2_s$, where R_j^2 is the estimated value of the *j*th

marker and $\widehat{\sigma}_y^2$ is the phenotypic variance of the *i*th trait; however, this is a biased estimator of $\sigma_{s_i}^2$ (Hospital et al. 1997).

Charcosset and Gallais (1996) and Hospital et al. (1997) proposed an unbiased estimator of $\sigma_{s_i}^2$ based on all the selected markers using the adjusted coefficient of multiple determination, i.e.,

$$R_{Adj}^2 = 1 - \frac{n-1}{n-M-1}\left(1-R^2\right) = \frac{\widehat{\sigma}_s^2}{\widehat{\sigma}_y^2}, \tag{4.32b}$$

whence we can obtain a unbiased estimator of $\sigma_{s_i}^2$ as $\widehat{\sigma}_y^2 R_{Adj}^2 = \widehat{\widehat{\sigma}}_{s_i}^2$ by jointly using all the markers that affect the phenotypic values. The problem with Eq. (4.32b) is that the R_{Adj}^2 values could be negative; in that case, the estimated value of $\sigma_{s_i}^2$ would also be negative. One additional problem with Eq. (4.32b) is that the R_{Adj}^2 values can produce $\widehat{\sigma}_s^2$ values that are higher than those of the estimated variance of the breeding values $\widehat{\sigma}_g^2$.

Using Eqs. (4.32a) and (4.32b), we can estimate $\sigma_{s_i}^2$, but from them it is not clear how we can estimate the covariance between two different estimated marker score values.

Consider the case of the PHT values described in Sect. 4.3.1 of this chapter, where $M = 7$, $n = 247$, and the estimated variance of PHT was $\widehat{\sigma}_{PHT}^2 = 191.81$. The estimated values of R^2 for each of the seven markers were 0.0038, 0.0005, 0.006, 0.0013, 0.0036, 0.0114, and 0.0298, whence, by multiplying each estimated R^2 value by $\widehat{\sigma}_{PHT}^2 = 191.81$ and summing the results, we found that the estimated value of $\sigma_{s_{PHT}}^2$ was $\widehat{\widehat{\sigma}}_{s_{PHT}}^2 = 9.78$. In this case, the estimated portion of the genetic variance attributable to $\widehat{\widehat{\sigma}}_{s_{PHT}}^2 = 9.78$ was $\widehat{q}_{PHT} = \dfrac{9.78}{83} = 0.1178$; thus, when we estimated $\sigma_{s_{PHT}}^2$ according to Eq. (4.32a), the seven markers explained only 11.78% of the genetic variance associated with PHT.

The estimated value of R_{Adj}^2 for the seven markers jointly was 0.06, whence $\widehat{\sigma}_{s_{PHT}}^2 = (191.81)(0.06) = 11.50$ is an estimate of $\sigma_{s_{PHT}}^2$. In the latter case, the estimated portion of the genetic variance attributable to $\widehat{\sigma}_{s_{PHT}}^2 = 11.50$ was $\widehat{q}_{PHT} = \dfrac{11.5}{83} = 0.1385$; that is, according to Eq. (4.32b), the seven markers explain 13.85% of the genetic variance associated with PHT.

One additional way of estimating the variance of the marker score $\sigma_{s_i}^2$ was proposed by Lange and Whittaker (2001) as

$$\frac{1}{n-1}\sum_{i=1}^{n}\left(\widehat{s}_i - \widehat{\mu}_{s_i}\right)^2, \tag{4.33}$$

where $\widehat{s}_i = \sum_{j=1}^{M} \widehat{\theta}_j x_j$ and $\widehat{\mu}_{s_i}$ is the mean of $\widehat{s}_i$ values. The covariance between the ith and jth marker scores can be estimated as the cross products of the marker score values divided by $n - 1$. Note that in this case, the number of markers associated with the ith and jth traits may be different.

For the PHT values described in Sect. 4.3.1 of this chapter, where $n = 247$, the estimated value of $\sigma^2_{s_i}$ was $\widehat{\sigma}^2_{s_{PHT}} = 15.75$ and the estimated portion of the genetic variance attributable to $\widehat{\sigma}^2_{s_{PHT}} = 15.75$ was $\widehat{q}_{PHT} = \dfrac{15.75}{83} = 0.1897$. That is, the seven markers jointly explain 18.97% of the genetic variance associated with PHT according to Eq. (4.33).

4.3.3 Estimating LMSI Selection Response and Efficiency

With the estimated phenotypic variances ($\widehat{\sigma}^2_{PHT} = 191.81$), the estimated genetic variance ($\widehat{\sigma}^2_{g_{PHT}} = 83.0$) and the estimated marker score variances: $\widehat{\sigma}^2_{\widehat{s}_{PHT}} = 48.23$ (Eq. 4.29), $\widehat{\sigma}^2_{\widehat{s}_{PHT}} = 9.78$ (Eq. 4.32a), $\widehat{\sigma}^2_{s_{PHT}} = 11.50$ (Eq. 4.32b), and $\widehat{\sigma}^2_{s_{PHT}} = 15.75$ (Eq. 4.33), we can estimate the LMSI coefficient, selection response, and efficiency.

Using the estimated value $\widehat{\sigma}^2_{\widehat{s}_{PHT}} = 48.23$ obtained with Eq. (4.29), it is possible to estimate the LMSI weight as $\widehat{\beta}_{PHT} = \dfrac{\widehat{\sigma}^2_{g_{PHT}} - \widehat{\sigma}^2_{s_{PHT}}}{\widehat{\sigma}^2_{PHT} - \widehat{\sigma}^2_{s_{PHT}}} = \dfrac{83.0 - 48.23}{191.81 - 48.23} = 0.242$, whereas for $\widehat{\sigma}^2_{\widehat{s}_{PHT}} = 9.78$, $\widehat{\sigma}^2_{s_{PHT}} = 11.50$, and $\widehat{\sigma}^2_{s_{PHT}} = 15.75$, the estimated values of β_{PHT} were 0.402, 0.40, and 0.382 respectively. The latter results indicate that the estimated values of β_{PHT} associated with the phenotypic values tend to decrease when the estimated values of the variance of the marker score increase. This means that at the limit, when all the genetic variance is explained by the markers, the estimated values of β_{PHT} are zero and the estimated LMSI is equal to $\widehat{I}_M = \widehat{s}$. Thus, for trait PHT, when the estimated values of β_{PHT} are not zero, the estimated LMSI can be written as $\widehat{I}_{M_{PHT}} = \widehat{s}_{PHT} + \widehat{\beta}_{PHT}\left(PHT_i - \widehat{s}_{PHT}\right)$. The $\widehat{I}_{M_{PHT}}$ values are used to predict, rank, and select the net genetic merit value of each individual candidate for selection.

Based on the result $\widehat{\sigma}^2_{\widehat{s}_{PHT}} = 48.23$ obtained with Eq. (4.29) and using a selection intensity of 10% (k_I= 1.755), the estimated LMSI selection response can be obtained as

$$\begin{aligned}\widehat{R}_M &= k_I\sqrt{\frac{\widehat{\sigma}_g^2\left(\widehat{\sigma}_g^2-\widehat{\sigma}_s^2\right)+\widehat{\sigma}_s^2\left(\widehat{\sigma}_y^2-\widehat{\sigma}_g^2\right)}{\widehat{\sigma}_y^2-\widehat{\sigma}_s^2}}\\ &= 1.755\sqrt{\frac{83(83-48.23)+48.23(191.81-83)}{191.81-48.23}}\\ &= 1.755\sqrt{56.65}=13.21.\end{aligned}$$

In a similar manner, using the result $\widehat{\sigma}_{s_{PHT}}^2 = 15.75$, the estimated selection response was $\widehat{R}_M = 1.755\sqrt{\frac{83(83-15.75)+15.75(191.81-83)}{191.81-15.75}} = 1.755\sqrt{41.44}$ $= 11.30$. With $\widehat{\sigma}_{s_{PHT}}^2 = 9.78$ and $\widehat{\sigma}_{s_{PHT}}^2 = 11.50$, the estimated values of the LMSI selection responses were 10.99 and 11.10 respectively. The latter results indicate that the estimated values of the LMSI selection responses tend to increase when the estimated values of the variance of the marker score increase.

We can estimate LMSI versus phenotypic efficiency for one trait as $\widehat{\lambda}_M = \sqrt{\frac{\widehat{q}}{\widehat{h}^2}+\frac{\left(1-\widehat{q}\right)^2}{1-\widehat{q}\widehat{h}^2}}$, where $\widehat{h}^2$ is the estimated trait heritability and $\widehat{q} = \frac{\widehat{\sigma}_s^2}{\widehat{\sigma}_g^2}$ is the estimated portion of additive genetic variance explained by the markers. When $\widehat{\sigma}_{s_{PHT}}^2 = 48.23$, $\widehat{q}_{PHT} = \frac{48.23}{83} = 0.5811$, and $\widehat{h}^2 = 0.433$, the estimated LMSI efficiency was $\widehat{\lambda}_M = \sqrt{1.58} = 1.25$. For $\widehat{\sigma}_{s_{PHT}}^2 = 15.75$, $\widehat{\sigma}_{s_{PHT}}^2 = 9.78$, and $\widehat{\sigma}_{s_{PHT}}^2 = 11.50$, the estimated portions of the additive genetic variance explained by the markers were $\widehat{q}_{PHT} = \frac{15.75}{83} = 0.1897$, $\widehat{q}_{PHT} = \frac{9.78}{83} = 0.1178$, and $\widehat{q}_{PHT} = \frac{11.5}{83} = 0.1385$ respectively, whence the estimated LMSI efficiencies were 1.1, 1.04, and 1.05 respectively. The latter results indicate that the estimated values of LMSI efficiency tend to increase when the estimated values of the variance of the marker score increase (Fig. 4.1).

Figure 4.1 presents the change in LMSI efficiency with respect to phenotypic selection for different values of the variance of the marker score when the phenotypic (191.81) and genetic (83) variances are fixed. In a similar manner, Fig. 4.2 presents the change in the LMSI selection response for different values of the variance of the marker score when the phenotypic (191.81) and genetic (83) variances are fixed. In effect, LMSI efficiency and the selection response depend on the genetic variance explained by the markers.

4.3.4 *Estimating the Variance of the Marker Score in the Multi-Trait Case*

Equation (4.33) can be used in the multi-trait context when the numbers of markers associated with the *i*th and *j*th traits are different. Also, it is possible to adapt Eqs. (4.32a) and (4.32b) to the multi-trait case. However, in the latter case, in addition to the markers linked to the QTL that affect one specific trait, we need to find markers that affect more than one trait, which may be very difficult. For this reason, in the multi-trait context, Eqs. (4.32a) and (4.32b) could be used to estimate the variance of the marker score (**S**) without preselecting the markers that affect the phenotypic traits, only when the number of genotypes is higher than the number of markers.

Let $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_r$ be r independent multivariate normal vectors of observations, each with n observations, such that $\mathbf{Y} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1t} \\ y_{21} & y_{22} & \cdots & y_{2t} \\ \vdots & \vdots & \cdots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nt} \end{bmatrix}$ is a matrix $n \times t$ of observations for t traits; then, the multivariate linear regression model can be written as $\mathbf{Y} = \mathbf{XB} + \mathbf{U}$, where $\mathbf{X}$ is a matrix $n \times m$ (m= number of markers and $m < n$) of known coded marker values, $\mathbf{B}$ is a matrix $m \times n$ of regression coefficients, and $\mathbf{U}$ is a matrix $n \times t$ of unobserved random disturbance whose rows for given $\mathbf{X}$ are uncorrelated, each with mean $\mathbf{0}$ and common covariance matrix $\mathbf{E}$ (Mardia et al. 1982; Rencher 2002). According to the least squares method of estimation, $\widehat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ is an estimator of $\mathbf{B}$ and $\widehat{\mathbf{E}} = \dfrac{\left(\mathbf{Y} - \widehat{\mathbf{B}}\mathbf{X}\right)'\left(\mathbf{Y} - \widehat{\mathbf{B}}\mathbf{X}\right)}{n - m - 1}$ is an estimator of the residual covariance matrix $\mathbf{E}$ assuming that $n > m$ (Johnson and Wichern 2007).

Note that $1 - R^2 = \dfrac{\widehat{\mathbf{e}}'\widehat{\mathbf{e}}}{\mathbf{y}'\mathbf{y}}$, where $\widehat{\mathbf{e}}$ is a vector of estimated residual values of the model $y_i = \theta_0 + \sum_{j \in M} \theta_j x_j + e_i$ and R^2 is the coefficient of multiple determination (Eq. 4.32a). In addition, as in the multi-trait context the estimated matrix of residuals is $\widehat{\mathbf{U}} = \mathbf{Y} - \widehat{\mathbf{B}}\mathbf{X}$, $1 - R^2$ can be written as $\mathbf{D} = (\mathbf{Y}'\mathbf{Y})^{-1}\widehat{\mathbf{U}}'\widehat{\mathbf{U}}$ (Mardia et al. 1982), whence R^2 in the multivariate context can written as

$$\mathbf{R}^2 = \mathbf{I} - \mathbf{D} = \widehat{\mathbf{P}}^{-1}\widehat{\mathbf{S}}, \tag{4.34a}$$

whereas R^2_{Adj} (Eq. 4.32b) can be written as

$$\mathbf{R}^2_{Adj} = \mathbf{I} - \frac{n-1}{n-m-1}\mathbf{D} = \widehat{\mathbf{P}}^{-1}\widehat{\mathbf{S}}, \tag{4.34b}$$

where $\mathbf{I}$ is an identity matrix $t \times t$, $\widehat{\mathbf{P}}^{-1}$ is the inverse of the estimated covariance matrix of phenotypic values ($\widehat{\mathbf{P}}$), and $\widehat{\mathbf{S}}$ is the estimated covariance matrix of marker score values. From Eq. (4.34b),

$$\widehat{\mathbf{P}}\mathbf{R}^2_{Adj} = \widehat{\mathbf{S}} \tag{4.34c}$$

is an unbiased estimator of matrix $\widehat{\mathbf{S}}$, whereas $\widehat{\mathbf{P}}\mathbf{R}^2 = \widehat{\mathbf{S}}$ (Eq. 4.34a) is a biased estimator of matrix $\widehat{\mathbf{S}}$. The main problem of Eq. (4.34c) is that the diagonal elements of $\widehat{\mathbf{S}}$ could be negative.

From the maize F_2 population including 247 genotypes (each one with two repetitions) and 195 molecular markers described in Sect. 4.3.1, we used two traits—PHT (cm) and EHT (cm)—to illustrate the multivariate method of estimating the LMSI parameters. The estimated phenotypic and genetic covariance matrices were $\widehat{\mathbf{P}} = \begin{bmatrix} 191.81 & 106.89 \\ 106.89 & 167.93 \end{bmatrix}$ and $\widehat{\mathbf{C}} = \begin{bmatrix} 83.00 & 57.44 \\ 57.44 & 59.80 \end{bmatrix}$, whereas the estimated covariance matrix of marker scores, using Eq. (4.33), was $\widehat{\mathbf{S}} = \begin{bmatrix} 15.750 & 0.983 \\ 0.983 & 28.083 \end{bmatrix}$. When we used Eq. (4.34a) and Eq. (4.34c), we obtained estimated values of the variance and covariance of the marker scores that were higher than the genetic values (data not presented). Equations (4.29) and (4.31) are used later to compare LMSI efficiency versus GW-LMSI efficiency using the simulated data described in Chap. 2, Sect. 2.8.1.

With matrices $\widehat{\mathbf{P}}$, $\widehat{\mathbf{C}}$, and $\widehat{\mathbf{S}}$, and the vector of economic weights $\mathbf{a}' = [\mathbf{w}' \quad \mathbf{0}']$, where $\mathbf{w}' = [-1 \quad -1]$ and $\mathbf{0}' = [0 \quad 0]$, we obtained the estimated matrices $\widehat{\mathbf{T}} = \begin{bmatrix} \widehat{\mathbf{P}} & \widehat{\mathbf{S}} \\ \widehat{\mathbf{S}} & \widehat{\mathbf{S}} \end{bmatrix}$ and $\mathbf{Z} = \begin{bmatrix} \widehat{\mathbf{C}} & \widehat{\mathbf{S}} \\ \widehat{\mathbf{S}} & \widehat{\mathbf{S}} \end{bmatrix}$, whence the estimated LMSI vector of coefficients was $\widehat{\boldsymbol{\beta}}' = \mathbf{a}'\widehat{\mathbf{Z}}_M\widehat{\mathbf{T}}_M^{-1} = [-0.59 \quad -0.18 \quad -0.41 \quad -0.82]$. Using a selection intensity of 10% ($k_I = 1.755$), the estimated LMSI selection response and the expected genetic gains per trait were $\widehat{R}_M = k_I\sqrt{\widehat{\boldsymbol{\beta}}'\widehat{\mathbf{T}}_M\widehat{\boldsymbol{\beta}}} = 20.41$ and $\widehat{\mathbf{E}}'_M = k_I\dfrac{\widehat{\boldsymbol{\beta}}'\widehat{\mathbf{Z}}_M}{\sqrt{\widehat{\boldsymbol{\beta}}'\widehat{\mathbf{T}}_M\widehat{\boldsymbol{\beta}}}} = [-10.09 \quad -10.31 \quad -2.53 \quad -4.39]$ respectively, whereas the estimated LMSI accuracy was $\widehat{\rho}_{H\hat{I}_M} = \dfrac{\widehat{\sigma}_{I_M}}{\widehat{\sigma}_H} = 0.72$.

The estimated LPSI parameters (see Chap. 2 for details) using the phenotypic information from the maize F_2 population for traits PHT and EHT are as follows. The estimated LPSI vector of coefficients was $\widehat{\mathbf{b}}' = \mathbf{w}'\widehat{\mathbf{C}}\widehat{\mathbf{P}}^{-1} = [-0.53 \quad -0.36]$, and, with a selection intensity of 10% ($k_I = 1.755$), the estimated LPSI selection response and the expected genetic gains per trait were $\widehat{R}_I = k_I\sqrt{\widehat{\mathbf{b}}'\widehat{\mathbf{P}}\widehat{\mathbf{b}}} = 18.97$ and $\widehat{\mathbf{E}}' = k_I\dfrac{\widehat{\mathbf{b}}'\widehat{\mathbf{C}}}{\widehat{\sigma}_I} = [-10.52 \quad -8.45]$ respectively, whereas the estimated LPSI accuracy was $\widehat{\rho}_{H\hat{I}} = \dfrac{\widehat{\sigma}_I}{\widehat{\sigma}_H} = 0.67$.

We can determine LMSI efficiency versus LPSI efficiency to predict the net genetic merit using the ratio of estimated accuracy values $\widehat{\rho}_{H\hat{I}_M} = 0.72$ and $\widehat{\rho}_{H\hat{I}} = 0.67$ of the LMSI and LPSI respectively, i.e., $\widehat{\lambda}_M = \dfrac{0.72}{0.67} = 1.075$, whence, according to Eq. (4.19), the estimated LMSI efficiency versus the LPSI efficiency, in percentage terms, was $\widehat{p}_M = 100(1.075 - 1) = 7.5$. That is, for these data, the estimated LMSI efficiency was only 7.5% greater than LPSI efficiency at predicting the net genetic merit.

4.4 Estimating the GW-LMSI Parameters in the Asymptotic Context

Lange and Whittaker (2001) proposed the GW-LMSI. However, these authors did not provide detailed procedures for estimating matrices **P**, **C**, **W**, and **M**. They indicated that matrix **C** can be estimated using the estimated matrix of covariance of marker scores ($\widehat{\mathbf{S}}$) and that matrices **P**, **W**, and **M** can be estimated *directly by their empirical variances and covariances*, but this assertion does not indicate a clear method for estimating those covariance matrices. In Chap. 2, we described the REML method of estimating **C** and **P**. Crossa and Cerón-Rojas (2011) described matrices **W** and **M** in a doubled haploid population. In this study, we describe and estimate matrices **W** and **M** for an F_2 population in the asymptotic context according to the Wright and Mowers (1994) approach, which is based on regressing phenotype values on marker coded values. We used this latter approach to estimate **W** and **M**, because it is a clearer estimation method than that of Lange and Whittaker (2001); however, the Wright and Mowers (1994) approach is an asymptotic method and should be regarded with precaution.

Matrix **M** is the covariance matrix of the molecular marker code values. All marker information used to construct matrix **M** is presented in Table 4.2. Based on this information, we found that the expectations ($E(X_1)$ and $E(X_2)$) and the variances ($V(X_1)$ and $V(X_2)$) of the marker coded values X_1 and X_2 are $E(X_1) = E(X_2) = 0$ and $V(X_1) = V(X_2) = 1$, whereas the covariance ($Cov(X_1, X_2)$) and correlation ($Corr(X_1, X_2)$), between X_1 and X_2 were

$$Cov(X_1, X_2) = Corr(X_1, X_2) = 1 - 2\delta. \tag{4.35}$$

Thus, as the variances of X_1 and X_2 are equal to 1, the correlation between X_1 and X_2 is $Corr(X_1, X_2) = \frac{Cov(X_1, X_2)}{\sqrt{V(X_1)V(X_2)}} = 1 - 2\delta$, i.e., the covariance and correlation between X_1 and X_2 are the same. Equation (4.35) results indicate that if we perform the same operation with many markers, we will obtain similar results; they also indicate that this is the way to construct matrix **M**.

Table 4.2 Marker genotypes, expected frequency, and coded values (X_1 and X_2) of the marker genotypes in an F_2 population

Marker genotype	Expected frequency	X_1	X_2
A_1B_1/A_1B_1	$(1-\delta)^2/4$	1	1
A_1B_1/A_1B_2	$2(\delta-\delta^2)/4$	1	0
A_1B_2/A_1B_2	$\delta^2/4$	1	−1
A_1B_1/A_2B_1	$2(\delta-\delta^2)/4$	0	1
A_1B_2/A_2B_1	$2(1-2\delta+2\delta^2)/4$	0	0
A_1B_2/A_2B_2	$2(\delta-\delta^2)/4$	0	−1
A_2B_1/A_2B_1	$\delta^2/4$	−1	1
A_2B_1/A_2B_2	$2(\delta-\delta^2)/4$	−1	0
A_2B_2/A_2B_2	$(1-\delta)^2/4$	−1	−1

Let $\mathbf{X}$ be a matrix of coded markers of size $n \times m$, where $n \geq m$ and m= number of markers; then according to Wright and Mowers (1994), because all marker information is contained in matrix $\mathbf{X}'\mathbf{X}$, when the number of observations (n) tends to infinity, the product $\mathbf{x}_i'\mathbf{x}_j/n$ tends to the covariance between markers ith and jth, whence matrix $n^{-1}\mathbf{X}'\mathbf{X}$ should tend to the covariance matrix between the markers that conform matrix $\mathbf{X}$ with the ijth element equal to $(0.5 - \delta_{ij})$. Thus, matrix $2n^{-1}\mathbf{X}'\mathbf{X}$ should tend to a covariance matrix where the ijth entry is equal to $(1 - 2\delta_{ij})$. Based on the latter result, an estimator of matrix $\mathbf{M}$ in the asymptotic context is

$$\widehat{\mathbf{M}} = 2n^{-1}\mathbf{X}'\mathbf{X}. \tag{4.36}$$

Equation (4.36) is an asymptotic result and should be taken with caution. To date, there has been no clear method for estimating $\mathbf{M}$ in the non-asymptotic context; for this reason, Eq. (4.36) is used to estimate the GW-LMSI parameters.

Assume that a QTL is between the two markers in Table 4.2; then, δ can be written as $\delta = r_1 + r_2 - 2r_1r_2$, where r_1 and r_2 denote the recombination frequency between marker 1 and marker 2 respectively, with the QTL between them. When the number of genotypes or individuals tends to infinity, the covariance between the phenotypic trait values (y) and the marker 1 coded values (X_1) in an F_2 population can be written as

$$Cov(X_1, y) = \frac{1}{2}\alpha_1(1 - 2r_1), \tag{4.37}$$

where $\alpha_1(1 - 2r_1)$ is the portion of the additive effect (α_1) of the QTL linked to marker 1 (Edwards et al. 1987), and r_1 is the recombination frequency between the QTL and marker 1. We can assume that for many markers, the covariance of the phenotypic values is similar to Eq. (4.37), whence matrix $\mathbf{W}$ can be obtained.

Let $\mathbf{y}$ be a vector $n \times 1$ of recorded phenotypic values, where n denotes the number of observation or records, and $\mathbf{X}$ is a matrix of coded markers of size $n \times m$.

When n tends to infinity, $2n^{-1}\mathbf{X}'\mathbf{y}$ tends to be a vector with elements equal to $\alpha_i(1-2r_i)$, where α_i is the additive effect of the ith QTL linked to the ith marker, and r_i is the recombination frequency between the ith QTL and the ith marker. Now let $\mathbf{Y} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1t} \\ y_{21} & y_{22} & \cdots & y_{2t} \\ \vdots & \vdots & \cdots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nt} \end{bmatrix}$ be a matrix of observations for t traits; then, an estimator of matrix $\mathbf{W}$ in the asymptotic context is

$$\widehat{\mathbf{W}} = 2n^{-1}\mathbf{X}'\mathbf{Y}. \tag{4.38}$$

Once again, Eq. (4.38) is an asymptotic result and should be accepted with caution. But to date, there has been no clear method for estimating $\mathbf{W}$ in the non-asymptotic context; for this reason, Eq. (4.38) is used to estimate the GW-LMSI parameters.

4.5 Comparing LMSI Versus LPSI and GW-LMSI Efficiency

To compare LMSI efficiency versus GW-LMSI efficiency for predicting the net genetic merit, we use the simulated data set described in Chap. 2, Sect. 2.8.1.

Figure 4.4 presents the estimated accuracy values of the LPSI ($\widehat{\rho}_{H\hat{I}} = \frac{\widehat{\sigma}_{\widehat{I}}}{\widehat{\sigma}_H}$), the LMSI ($\widehat{\rho}_{H\hat{I}_M} = \frac{\widehat{\sigma}_{\widehat{I}_M}}{\widehat{\sigma}_H}$), and the GW-LMSI ($\widehat{\rho}_{H\hat{I}_W} = \frac{\widehat{\sigma}_{\widehat{I}_W}}{\widehat{\sigma}_H}$) for five simulated selection cycles. In addition, Table 4.3 presents the estimated LPSI, LMSI, and GW-LMSI selection responses, the estimated LPSI, LMSI, and GW-LMSI variances of the predicted error ($\left(1-\widehat{\rho}^2_{H\hat{I}}\right)\widehat{\sigma}^2_H$, $\left(1-\widehat{\rho}^2_{H\hat{I}_M}\right)\widehat{\sigma}^2_H$ and $\left(1-\widehat{\rho}^2_{H\hat{I}_W}\right)\widehat{\sigma}^2_H$ respectively), the ratios of the estimated LMSI accuracy to the estimated LPSI accuracy and the estimated LMSI accuracy to the estimated GW-LMSI accuracy, expressed as percentages (Eq. 4.19), for five simulated selection cycles.

According to Fig. 4.4, for this data set the estimated LMSI accuracy ($\widehat{\rho}_{H\hat{I}_M}$) was higher than the estimated LPSI and GW-LMSI accuracy ($\widehat{\rho}_{H\hat{I}}$ and $\widehat{\rho}_{H\hat{I}_W}$ respectively), for the five simulated selection cycles, that is, $\widehat{\rho}_{H\hat{I}_M} > \widehat{\rho}_{H\hat{I}} > \widehat{\rho}_{H\hat{I}_W}$. In a similar manner, Table 4.3 results indicate that the estimated LMSI selection response ($\widehat{R}_M$) was higher than the estimated LPSI and GW-LMSI selection responses ($\widehat{R}_I$ and $\widehat{R}_W$ respectively): $\widehat{R}_M > \widehat{R}_I > \widehat{R}_W$.

Note that the estimated LPSI, LMSI, and GW-LMSI variances of the predicted error, and the estimated LMSI efficiency versus LPSI efficiency and versus GW-LMSI efficiency (expressed in percentages) are related to the estimated

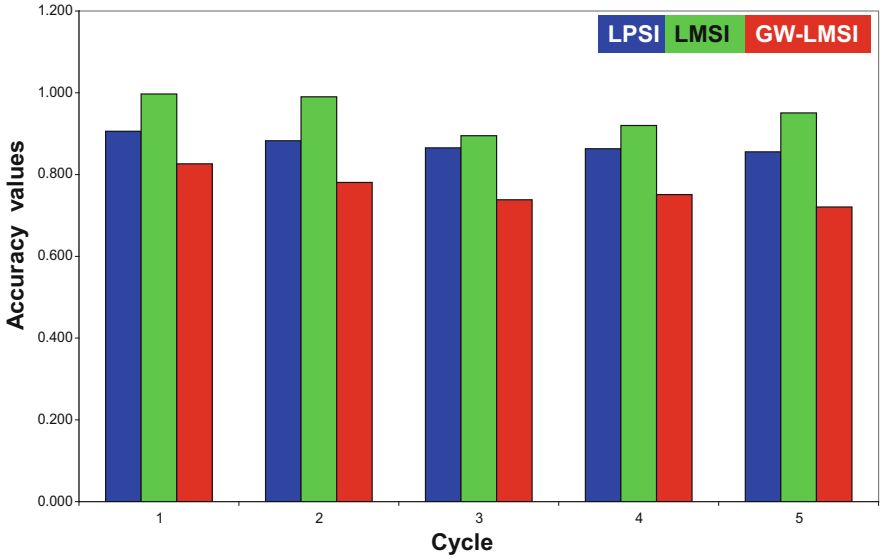

Fig. 4.4 Estimated correlation values of the linear phenotypic selection index (LPSI), the linear molecular selection index (LMSI), and the genome-wide LMSI (GW-LMSI) with the net genetic merit for four traits, 2500 markers and 500 genotypes (each with four repetitions) in one environment for five simulated selection cycles

Table 4.3 Estimated linear phenotypic, molecular, and genome-wide selection indices (LPSI, LMSI, and GW-LMSI respectively), selection responses and variance of the predicted error, and estimated ratio of LMSI accuracy to LPSI and GW-LMSI accuracy expressed in percentages for 4 traits, 2500 markers and 500 genotypes (each with four repetitions) in one environment for five simulated selection cycles

	Selection response			Variance of the predicted error			Efficiency of LMSI versus	
Cycle	LPSI	LMSI	GW-LMSI	LPSI	LMSI	GW-LMSI	LPSI	GW-LMSI
1	17.84	19.60	16.24	22.53	0.07	39.84	10.07	20.67
2	15.66	24.36	13.88	22.66	0.07	40.06	12.14	26.81
3	14.44	14.70	12.13	21.95	1.86	39.86	3.43	21.27
4	14.29	15.29	12.48	22.84	1.46	39.09	6.57	22.50
5	13.86	15.15	11.49	22.13	0.88	39.65	11.11	31.88
Average	15.22	17.82	13.24	22.42	0.87	39.70	8.66	24.63

LMSI, LPSI, and GW-LMSI accuracies, and that in all five selection cycles, $\widehat{\rho}_{H\hat{I}_M} > \widehat{\rho}_{H\hat{I}} > \widehat{\rho}_{H\hat{I}_W}$. This implies that the estimated LMSI variance of the predicted error was lower than the estimated LPSI and GW-LMSI variance of the predicted error. In a similar manner, because $\widehat{\rho}_{H\hat{I}_M} > \widehat{\rho}_{H\hat{I}} > \widehat{\rho}_{H\hat{I}_W}$, the estimated LMSI efficiency was higher than the estimated LPSI efficiency and the estimated GW-LMSI efficiency.

Based on Fig. 4.4 and Table 4.3 results, we conclude that the LMSI was a better predictor of the net genetic merit than the LPSI, and that the LPSI is a better predictor of the net genetic merit than the GW-LMSI for this simulated data set.

References

Bulmer MG (1980) The mathematical theory of quantitative genetics. Lectures in biomathematics. University of Oxford, Clarendon Press, Oxford

Charcosset A, Gallais A (1996) Estimation of the contribution of quantitative trait loci (QTL) to the variance of a quantitative trait by means of genetic markers. Theor Appl Genet 93:1193–1201

Crossa J, Cerón-Rojas JJ (2011) Multi-trait multi-environment genome-wide molecular marker selection indices. J Indian Soc Agric Stat 62(2):125–142

Dekkers JCM, Settar P (2004) Long-term selection with known quantitative trait loci. Plant Breed Rev 24:311–335

Edwards MD, Stuber CW, Wendel JF (1987) Molecular-marker-facilitated investigations of quantitative-trait loci in maize. I. Numbers, genomic distribution and types of gene action. Genetics 116:113–125

Hospital F, Moreau L, Lacoudre F, Charcosset A, Gallais A (1997) More on the efficiency of marker-assisted selection. Theor Appl Genet 95:1181–1189

Johnson RA, Wichern DW (2007) Applied multivariate statistical analysis, 6th edn. Pearson Prentice Hall, Upper Saddle River, NJ

Knapp SJ (1998) Marker-assisted selection as a strategy for increasing the probability of selecting superior genotypes. Crop Sci 38:1164–1174

Lande R, Thompson R (1990) Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics 124:743–756

Lange C, Whittaker JC (2001) On prediction of genetic values in marker-assisted selection. Genetics 159:1375–1381

Mardia KV, Kent JT, Bibby JM (1982) Multivariate analysis. Academic Press, New York

Moreau L, Charcosset A, Hospital F, Gallais A (1998) Marker-assisted selection efficiency in populations of finite size. Genetics 148:1353–1365

Moreau L, Hospital F, Whittaker J (2007) Marker-assisted selection and introgression. In: Balding DJ, Bishop M, Cannings C (eds) Handbook of statistical genetics, vol 1, 3rd edn. Wiley, New York, pp 718–751

Rencher AC (2002) Methods of multivariate analysis. Wiley, New York

Searle S, Casella G, McCulloch CE (2006) Variance components. Wiley, Hoboken, NJ

Whittaker JC (2003) Marker-assisted selection and introgression. In: Balding DJ, Bishop M, Cannings C (eds) Handbook of statistical genetics, vol 1, 2nd edn. Wiley, New York, pp 554–574

Wright AJ, Mowers RP (1994) Multiple regression for molecular marker, quantitative trait data from large F_2 population. Theor Appl Genet 89:305–312

Zhang W, Smith C (1992) Computer simulation of marker-assisted selection utilizing linkage disequilibrium. Theor Appl Genet 83:813–820

Zhang W, Smith C (1993) Simulation of marker-assisted selection utilizing linkage disequilibrium: the effects of several additional factors. Theor Appl Genet 86:492–496