

國立台灣大學生物資源暨農學院農藝學研究所生物統計組

碩士論文

Division of Biometry, Graduate Institute of Agronomy

College of Bioresources and Agriculture

National Taiwan University

Master Thesis

透過多批次外表型收集策略選拔優良基因型

A Sequential Batch Phenotyping Strategy for Detecting Superior
Genotypes

杜鎮宇

Zhen-Yu Tu

指導教授：廖振鐸 博士、蔡欣甫 博士

Advisor: Chen-Tuo Liao Ph.D, Shin-Fu Tsai Ph.D

中華民國 109 年 7 月

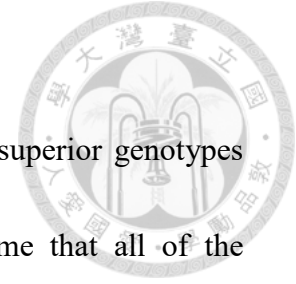
July, 2020

摘要

使用多批次外表型收集策略來從一組候選族群中找到一群優良基因型群體，以達到能節省外表型資料的蒐集，進而找到優良基因型群體。在本研究中我們假設候選族群已擁有基因型資料，並分多次選取部分個體收集其外表型，而後使用具有外表型資料的個體建立 GBLUP 多性狀模型、並估計候選族群個體的基因型值、對於不同性狀給予不同權重後相加成一個選拔指標，並進行排序。其中用於選取訓練族群個體的方法有 r-score、M-PGV、EI-PGV 以及 EI-PGV-fwd，而所有方法的第一組起始個體選取皆使用 r-score 的方法，因為 r-score 只需要使用基因型的資訊而不需要考慮外表型的資訊。多性狀模型的應用讓我們同時針對多個性狀進行估計、然後根據不同性狀的重要性進行加權總合、最終得到的值稱作 composite selection index (CSI)。針對排序後的 CSI 則使用 correctly identified proportion (CIP) 以及 normalized discounted cumulative gain (NDCG) 作為評估指標，這兩項指標可以對感興趣的前幾名個體進行評量，且 NDCG 還多考慮了排序的正確性。經由上述的流程，最終能夠輔助我們選拔出個體來進行外表型資料蒐集、使得模型有良好的估計與排序，進而有效率的找到優良的基因型群體。

關鍵字：多批次外表型收集策略、多性狀、選拔指標、r-score、GBLUP

Abstract



A sequential phenotyping strategy is proposed to detect a set of superior genotypes efficiently from a candidate population. In this study, we assume that all of the individuals in the candidate population have been already genotyped. The iterative searching process is composed of the following steps. Step 0: a starting training set is determined from the candidate population according to the r-score algorithm. Step 1: a multiple-trait GBLUP model is trained using the phenotype and genotype data of the current training set. Step 2: a composite selection index (CSI) is constructed and estimated for each individual in the candidate population with genotypes based on the resulting multiple-trait GBLUP model. Step 3: two assessment indices, correctly identified proportion (CIP) and normalized discounted cumulative gain (NDCG) are calculated based on the estimates of CSI for a set of candidate individuals, and are used to evaluate the accuracy for the detection of the superior individuals. Step 4: four acquisition functions, r-score, M-PGV, EI-PGV and EI-PGV-fwd, are used to select additional training set added with the current training set. We further provide a stopping rule for the sequential strategy for practical applications. Three genome datasets are analyzed to illustrate our proposed sequential phenotyping strategy.

Keywords: Sequential phenotyping strategy, Multiple traits, Composite Selection index, r-score, GBLUP

Contents



摘要	i
Abstract	ii
Introduction	1
Materials and Methods	5
44k Rice Dataset	5
Tropical Rice Breeding Lines Dataset	5
Wheat Dataset	6
Standardized Multiple-trait GBLUP Model	6
Composite Selection Index	7
The Distribution of Predicted Genotypic Values	8
The Expected Improvement Criteria for the CSI	9
r-score method	11
The Assessment Indices	12
Iterative Strategy	14
Criteria Comparison Based on Real Datasets	15
The Stopping Rule for the Iterative Strategy	16
Results	17
Criteria Comparison Based on Assessment Indices	17
The Stopping Rule for the Iterative Strategy	18
The True Genotypic Values for Specific Batches	19
Discussion	21
Reference	24
Tables and Figures	28

List of Figures



Figure 1: The assessment indices for case (i)	35
Figure 2: The assessment indices for case (ii)	38
Figure 3: The assessment indices for case (iii)	42
Figure 4: The assessment indices for case (iv)	46
Figure 5: The assessment indices for case (v)	50
Figure 6: The stopping rule for case (i)	51
Figure 7: The stopping rule for case (i)	52
Figure 8: The stopping rule for case (i)	54
Figure 9: The stopping rule for case (i)	56
Figure 10: The stopping rule for case (i)	58

List of Tables



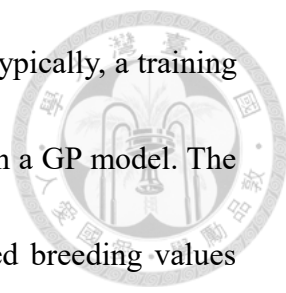
Table 1: The specified weight for trait i of each dataset	28
Table 2: The true genotypic values for case (i)	29
Table 3: The true genotypic values for case (ii)	30
Table 4: The true genotypic values for case (iii)	31
Table 5: The true genotypic values for case (iv)	32
Table 6: The true genotypic values for case (v)	33
Table 7: Comparison of A and B	34

Introduction



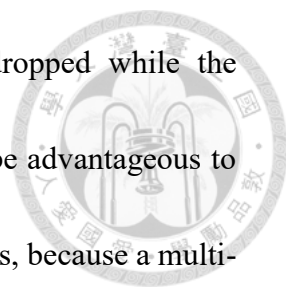
Food security issues have become very important since the rapid growth of the global population in the last few decades. Many innovative biotechnologies and breeding strategies have been applied to plant breeding for improving the yield and quality of crops (Tester and Langrude, 2010; Khoury et al., 2014). Although it has indeed achieved a remarkable improvement in the breeding selection process, the genetic diversity of crops has been gradually decreasing (Reif et al., 2005; Hyten et al., 2006). Genetic diversity is related to the potential of genetic improvement, further influencing the efficiency of breeding. Therefore, introgression of rich variation from wild, exotic, or indigenous germplasms becomes essential to promote the use of genetic diversity, and to enhance the efficiency of plant breeding programs (Tanksley and McCouch, 1997; McCouch et al., 2013). To tackle this problem, plant breeders first need to identify superior accessions from the germplasm collections. In this thesis, we focus on the identification of superior genotypes from a candidate population through a sequential phenotyping strategy. The proposed strategy is developed based on genomic prediction (GP), which can potentially accelerate the rate of genetic gain in crops.

The GP takes advantage of high-density DNA markers over a whole genome to predict the genotypic values, and then applies the estimated genotypic values to genomic selection in plant breeding (Meuwissen et al. 2001). The most common DNA



markers used in GP are single nucleotide polymorphisms (SNPs). Typically, a training population with known genotype and phenotype data is used to train a GP model. The resulting GP model is then employed to predict genomic estimated breeding values (GEBVs) for the individuals of a breeding population with known genotype data. The GP allows us to use limited phenotypic data to evaluate a large number of individuals with genotypes in the breeding population. The GP has been implemented for the two common objectives: (i) identify inbred lines either for hybrid parent development or cultivar release; (ii) increase the frequency of favorable alleles through rapid recurrent genomic selection (Gaynor et al. 2017).

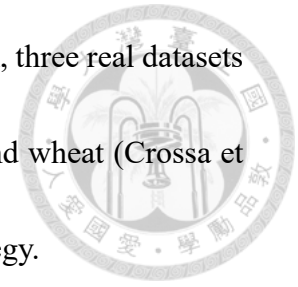
Mixed effects model methods have been widely used to GP such as ridge regression best linear unbiased predictor (rr-BLUP) model (Meuwissen et al. 2001), and genomic BLUP (GBLUP) model (VanRaden 2008). Specifically, GBLUP model can be extended to predict GEBVs for multiple traits simultaneously (Covarrubias-Pazaran, 2016). Moreover, Jia and Jannink (2012), Hayashi and Iwata (2013) and Guo et al. (2014) highlighted that multiple-trait GP models can provide better prediction accuracy than single-trait GP models for those traits with low heritabilities but highly correlated to the traits with high heritabilities. However, to evaluate the comprehensive performance of an individual under multiple traits, a suitable selection index is required to identify superior ones from a candidate population (Schulthess et al. 2016).



In the genomic era, genotyping costs have dramatically dropped while the phenotyping costs stay relatively constant. In this sense, it would be advantageous to sample individuals for selective phenotyping in more than one stages, because a multi-stage sampling scheme can reduce the size of the training population set in GP, hence the cost of phenotyping. Recently, Tanaka and Iwata (2018) proposed a multi-stage strategy using GP in pre-breeding to discover the best genotype from a candidate population. They implemented the concept of Bayesian optimization in the GP. The main idea of Bayesian optimization is to treat the desired objective function as a random variable, which is usually assumed to be a Gaussian process. Then an acquisition function, such as expected improvement (EI) or upper confidence bound, is constructed based on the posterior estimation for determining new query points to evaluate the objective function. The choice of the new query points should balance the trade-off between exploration and exploitation so that one can optimize the objective function using as few query points as possible (Shahriari et al. 2016; Gong et al. 2019).

In this thesis, we modify the strategy proposed by Tanaka and Iwata (2018) to identify superior individuals for multiple traits. We propose a new standardized multiple-trait GBLUP model to predict a composite selection index of multiple traits. Then, we implement the EI criterion to sample potential candidate individuals. Two indices of correctly identified proportion (CIP) and normalized discounted cumulative

gain (NDCG) are used to evaluate the proposed strategy. In addition, three real datasets of 44k rice (Zhao et al. 2011), tropical rice (Spindle et al. 2015) and wheat (Crossa et al. 2010) are analyzed to illustrate the sequential phenotyping strategy.



Materials and Methods



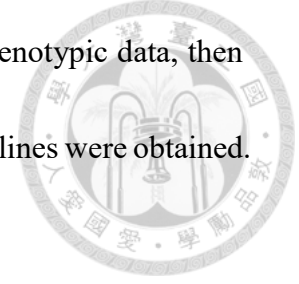
44k rice dataset

There are 413 rice accessions in the dataset, presented in Zhao et al. (2011), which has 36 traits in the phenotype data and 44100 SNP markers in the genotype data. Quality control has been already performed and 36901 SNP markers were retained with call rate $> 70\%$ and minor allele frequency > 0.01 , then impute the major allele to all missing position in genotype data. Here, we select two traits, flowering time at Arkansas (FTAA) and plant height (PLHE) for analyzing in this study. Since it should not have any missing data in phenotype data when performing the sequential strategy, we remove the accession which phenotype data are missing either in FTAA or PLHE. Finally, we have 36901 SNP markers, two traits without any missing and 373 accessions consists of 12 *aromatic*, 55 *aus*, 72 *indica*, 86 *temperate japonica*, 90 *tropical japonica*, and 58 admixed.

Tropical rice breeding lines dataset

We use a tropical rice breeding lines dataset which was presented in Spindel et al. (2015). It contains 363 lines and 73147 SNP markers for its genotype data. There are three traits in the dataset: yield (YLD), plant height (PH) and flowering time (FT). Since these data were collected from different years and seasons, these data have been already

adjusted by fitting a linear model. Integrating and averaging all phenotypic data, then merge the genotypic and phenotypic data, and finally 328 out of 363 lines were obtained.



Wheat dataset

The dataset was used in Crossa et al. (2010), which contains 599 accessions with grain yield data derived from four different environmental conditions in the phenotype data, and there are 1279 DArT markers in the genotype data.

Standardized Multiple-trait GBLUP Model

Let $\mathbf{w}_i = (\mathbf{y}_i - \bar{y}_i \mathbf{1}_n) / s_i$, where \bar{y}_i and s_i are the sample mean and the sample standard deviation of phenotypic values for trait i , i.e. $\mathbf{y}_i = [y_{i1}, \dots, y_{in}]^T$, for $i = 1, 2, \dots, t$. Also, let

$$\mathbf{w}_c = \begin{bmatrix} \mathbf{w}_1 \\ \vdots \\ \mathbf{w}_t \end{bmatrix}; \boldsymbol{\mu}_c = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_t \end{bmatrix}; \mathbf{g}_c = \begin{bmatrix} \mathbf{g}_1 \\ \vdots \\ \mathbf{g}_t \end{bmatrix} \text{ and } \mathbf{e}_c = \begin{bmatrix} \mathbf{e}_1 \\ \vdots \\ \mathbf{e}_t \end{bmatrix},$$

where μ_i , \mathbf{g}_i and \mathbf{e}_i denote the general mean, the vector of genotypic values and the vector of random errors for trait i , respectively. Then we consider the following standardized multiple-trait GBLUP model

$$\mathbf{w}_c = \boldsymbol{\mu}_c \otimes \mathbf{1}_n + \mathbf{g}_c + \mathbf{e}_c, \quad (1)$$

where $\mathbf{1}_n$ is the unit vector of order n and \otimes denotes the Kronecker product (Searle, 1982, P266). It is assumed that

$$\mathbf{g}_c \sim MVN(\mathbf{0}, \boldsymbol{\Sigma}_g \otimes \mathbf{K})$$

and

$$\mathbf{e}_c \sim MVN(\mathbf{0}, \boldsymbol{\Sigma}_e \otimes \mathbf{I}_n),$$



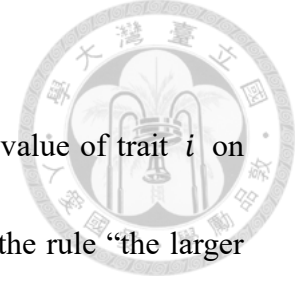
where $\mathbf{0}$ is the zero vector; \mathbf{K} is a genomic relationship matrix; $\boldsymbol{\Sigma}_g$ is the genetic variance-covariance matrix among traits; \mathbf{I}_n is the identity matrix of order n and $\boldsymbol{\Sigma}_e$ is the variance-covariance matrix of random errors among traits. Also, let

$$\boldsymbol{\Sigma}_g = \begin{bmatrix} \sigma_{g_1}^2 & \cdots & \sigma_{g_{1t}} \\ \vdots & \ddots & \vdots \\ \sigma_{g_{1t}} & \cdots & \sigma_{g_t}^2 \end{bmatrix} \text{ and } \boldsymbol{\Sigma}_e = \begin{bmatrix} \sigma_{e_1}^2 & \cdots & \sigma_{e_{1t}} \\ \vdots & \ddots & \vdots \\ \sigma_{e_{1t}} & \cdots & \sigma_{e_t}^2 \end{bmatrix}.$$

In this study, we use $\mathbf{K} = \mathbf{M}\mathbf{M}^T/p$, where \mathbf{M} is the standardized marker score matrix and p is the number of SNP markers. Let \mathbf{A} be the original marker score matrix with elements equal to -1, 0 and 1, corresponding to homozygous alleles (A_1A_1), heterozygous alleles (A_1A_2) and the other homozygous alleles (A_2A_2), respectively. Also, let m_{ij} and a_{ij} separately denote the $(ij)^{th}$ elements of \mathbf{M} and \mathbf{A} . Then, $m_{ij} = (a_{ij} - \bar{a}_j)/s_j$, where \bar{a}_j and s_j are the sample mean and the sample standard deviation of column j (corresponding to SNP j) in \mathbf{A} .

Composite Selection Index

To evaluate an individual with multiple traits, we propose a selection index simultaneously accounting for all the traits of interest. Define the composite selection index (CSI) for individual j as



$$CSI_{(j)} = \sum_{i=1}^t (\pm) p_i g_{ij},$$

where p_i is the specified weight for trait i ; g_{ij} is the genotypic value of trait i on individual j and the sign of (\pm) is taken “+” if the trait follows the rule “the larger the better”; otherwise “-” if the trait follows “the smaller the better”.

The Distribution of Predicted Genotypic Values

Let

$$\mathbf{w}_{c1} = \begin{bmatrix} \mathbf{w}_{11} \\ \vdots \\ \mathbf{w}_{t1} \end{bmatrix}; \mathbf{w}_{c2} = \begin{bmatrix} \mathbf{w}_{12} \\ \vdots \\ \mathbf{w}_{t2} \end{bmatrix}; \mathbf{g}_{c1} = \begin{bmatrix} \mathbf{g}_{11} \\ \vdots \\ \mathbf{g}_{t1} \end{bmatrix};$$

$$\mathbf{g}_{c2} = \begin{bmatrix} \mathbf{g}_{12} \\ \vdots \\ \mathbf{g}_{t2} \end{bmatrix}; \mathbf{e}_{c1} = \begin{bmatrix} \mathbf{e}_{11} \\ \vdots \\ \mathbf{e}_{t1} \end{bmatrix} \text{ and } \mathbf{e}_{c2} = \begin{bmatrix} \mathbf{e}_{12} \\ \vdots \\ \mathbf{e}_{t2} \end{bmatrix},$$

where \mathbf{w}_{i1} , \mathbf{g}_{i1} and \mathbf{e}_{i1} respectively denote the vectors of standardized phenotypic values, genotypic values and random errors for the training set. The training set is assumed to consist of n_1 individuals. Likewise, \mathbf{w}_{i2} , \mathbf{g}_{i2} and \mathbf{e}_{i2} denote the corresponding vectors for the remaining n_2 individuals not chosen in the training set (non-phenotyped set), where $n_1 + n_2 = n$. Thus, the standardized multiple-trait

GBLUP model of (1) can be equivalently written as

$$\begin{bmatrix} \mathbf{w}_{c1} \\ \mathbf{w}_{c2} \end{bmatrix} = \boldsymbol{\mu}_c \otimes \begin{bmatrix} \mathbf{1}_{n_1} \\ \mathbf{1}_{n_2} \end{bmatrix} + \begin{bmatrix} \mathbf{g}_{c1} \\ \mathbf{g}_{c2} \end{bmatrix} + \begin{bmatrix} \mathbf{e}_{c1} \\ \mathbf{e}_{c2} \end{bmatrix},$$

where

$$\begin{bmatrix} \mathbf{g}_{c1} \\ \mathbf{g}_{c2} \end{bmatrix} \sim MVN \left(\mathbf{0}, \boldsymbol{\Sigma}_g \otimes \begin{bmatrix} \mathbf{K}_{11} & \mathbf{K}_{12} \\ \mathbf{K}_{21} & \mathbf{K}_{22} \end{bmatrix} \right).$$

We used the R package *sommer* (Covarrubias-Pazaran, 2016) to find the REMLs for Σ_g and Σ_e ; and BLUP for g_{c1} using the training set data. That was based on the following model: $w_{c1} = \mu_c \otimes \mathbf{1}_{n1} + g_{c1} + e_{c1}$ where $g_{c1} \sim MVN(\mathbf{0}, \Sigma_g \otimes \mathbf{K}_{11})$.

These estimated values are denoted as $\hat{\mu}_c$, $\hat{\Sigma}_g$, $\hat{\Sigma}_e$ and \hat{g}_{c1} .

Under the condition that $\hat{\mu}_c$, $\hat{\Sigma}_g$, $\hat{\Sigma}_e$ and \hat{g}_{c1} are all assumed to be fixed and known values, the distribution of predicted genotypic values (PGVs) for the non-phenotyped set is given by

$$\tilde{g}_{c2} | (\hat{\mu}_c, \hat{\Sigma}_g, \hat{\Sigma}_e, \hat{g}_{c1}) \sim MVN(\hat{\mu}_{gc2}, \hat{\Sigma}_{gc2}), \quad (2)$$

where $\hat{\mu}_{gc2} = \hat{g}_{c2} = \mathbf{K}_{21}(\mathbf{K}_{11})^{-1}\hat{g}_{c1}$ and $\hat{\Sigma}_{gc2} = \hat{\Sigma}_g \otimes (\mathbf{K}_{22} - \mathbf{K}_{21}(\mathbf{K}_{11})^{-1}\mathbf{K}_{12})$.

Note that the calculation for $\hat{\mu}_{gc2}$ and $\hat{\Sigma}_{gc2}$ doesn't involve $\hat{\Sigma}_e$, i.e. the distribution of \hat{g}_{c2} is free from the random or environmental variation.

The Expected Improvement criteria for the CSI

Let $\tilde{g}_{c2(j)} = [\tilde{g}_{c21(j)} \cdots \tilde{g}_{c2t(j)}]^T$ be the vector of the t genotypic values on individual j in Expression (2). Then the $CSI_{(j)}$ for the individuals in the non-phenotyped set is given by

$$\widetilde{CSI}_{(j)} = \sum_{i=1}^t (\pm) p_i \tilde{g}_{c2i(j)}.$$

Clearly, $\widetilde{CSI}_{(j)}$ is a linear combination of $\tilde{g}_{c2(j)}$, so that its distribution can be easily obtained from Expression (2). The distribution of $\widetilde{CSI}_{(j)}$ is described as

$\widetilde{CSI}_{(j)} \sim N(\hat{\mu}_{CSI(j)}, \hat{\sigma}_{CSI(j)}^2)$. The improvement function for $\widetilde{CSI}_{(j)}$ is defined as

$$Im(\widetilde{CSI}_{(j)}) = \begin{cases} 0, & \text{if } \widetilde{CSI}_{(j)} < \hat{f}_M \\ \widetilde{CSI}_{(j)} - \hat{f}_M, & \text{otherwise,} \end{cases}$$

where \hat{f}_M is the maximal estimated $CSI_{(j)}$ value among the training set which is obtained from \hat{g}_{c1} . Here, $Im(\widetilde{CSI}_{(j)})$ is a random variable associated with the distribution of $\widetilde{CSI}_{(j)}$ and its expected value called the expected improvement (EI), can be derived as

$$EI(\widetilde{CSI}_{(j)}) = (\hat{\mu}_{CSI(j)} - \hat{f}_M) \Phi(Z_j) + \hat{\sigma}_{CSI(j)} \phi(Z_j), \quad (3)$$

where $Z_j = (\hat{\mu}_{CSI(j)} - \hat{f}_M) / \hat{\sigma}_{CSI(j)}$; $\Phi(\cdot)$ is the cumulative density function of the standard normal distribution and $\phi(\cdot)$ is the probability density function of the standard normal distribution.

Furthermore, let $\tilde{\mathbf{h}}$ denote the vector of $\widetilde{CSI}_{(j)}$ for the non-phenotyped set. The distribution of $\tilde{\mathbf{h}}$ can be denoted as $\tilde{\mathbf{h}} \sim MVN(\hat{\boldsymbol{\mu}}_{CSI}, \hat{\boldsymbol{\Sigma}}_{CSI})$. Partition $\tilde{\mathbf{h}}$, $\hat{\boldsymbol{\mu}}_{CSI}$ and $\hat{\boldsymbol{\Sigma}}_{CSI}$ as

$$\tilde{\mathbf{h}} = \begin{bmatrix} \widetilde{CSI}^* \\ \tilde{\mathbf{h}}^* \end{bmatrix}; \hat{\boldsymbol{\mu}}_{CSI} = \begin{bmatrix} \hat{\mu}_{CSI}^* \\ \hat{\boldsymbol{\mu}}^* \end{bmatrix} \text{ and } \hat{\boldsymbol{\Sigma}}_{CSI} = \begin{bmatrix} \hat{\sigma}_{CSI}^2 & \hat{\boldsymbol{\Sigma}}_{12}^* \\ \hat{\boldsymbol{\Sigma}}_{21}^* & \hat{\boldsymbol{\Sigma}}_{22}^* \end{bmatrix},$$

where \widetilde{CSI}^* represents the genotype with largest $EI(\widetilde{CSI}_{(j)})$ of Equation (3), i.e. the genotype with \widetilde{CSI}^* is the first selected from the non-phenotyped set. Subsequently, we searched for the next genotype with the largest $EI(\widetilde{CSI}_{(j)})$ among the remaining genotypes whose PGVs follow the conditional distribution

$$\tilde{\mathbf{h}}^* | (\widetilde{CSI}^* = \hat{\mu}_{CSI}^*) \sim MVN(\hat{\boldsymbol{\mu}}_{\tilde{\mathbf{h}}}, \hat{\boldsymbol{\Sigma}}_{\tilde{\mathbf{h}}}), \quad (4)$$

where $\hat{\boldsymbol{\mu}}_h^* = \hat{\boldsymbol{\mu}}^*$ and $\hat{\boldsymbol{\Sigma}}_h^* = \hat{\boldsymbol{\Sigma}}_{22}^* - \hat{\boldsymbol{\Sigma}}_{21}^* (\hat{\sigma}_{CSI}^2)^{-1} \hat{\boldsymbol{\Sigma}}_{12}^*$. Let $\tilde{h}_{(j)}^* \sim N(\hat{\mu}_{h(j)}^*, (\hat{\sigma}_{h(j)}^*)^2)$,

representing a marginal distribution in Expression (4), then the corresponding EI can

be derived as

$$EI(\tilde{h}_{(j)}^*) = (\hat{\mu}_{h(j)}^* - \hat{f}_M) \Phi(Z_{(j)}^*) + \hat{\sigma}_{h(j)}^* \phi(Z_{(j)}^*), \quad (5)$$

where $Z_{(j)}^* = (\hat{\mu}_{h(j)}^* - \hat{f}_M) / \hat{\sigma}_{h(j)}^*$. The EI of (3) is abbreviated as EI-PGV, and the EI

of (5) as EI-PGV-fwd. Also, M-PGV is the criterion using the mean values of PGVs.

r-score method

Ou and Liao (2019) proposed an optimization method to determine a training set for genomic selection. Their proposed method was derived from the Pearson's correlation between GEBVs and phenotypic values, called as r-score method. The r-score method was verified to be advantageous over some existing optimization method (Ou and Liao, 2019), and it can be used to choose an optimal training set from a candidate population with genotype data only.

In this study, we use r-score method to determine a starting training set for the sequential phenotyping strategy. The r-score criterion can be described as

$$r\text{-score} = \frac{q_{12}}{\sqrt{q_1 q_2}}$$

where

$$q_{12} = \text{Tr}[\mathbf{X}_0^T (\mathbf{I}_{n_0} - \bar{\mathbf{J}}_{n_0}) \mathbf{X}_0 \mathbf{A} \mathbf{X}];$$



$$q_1 = (n_0 - 1) + Tr[\mathbf{X}_0^T(\mathbf{I}_{n_0} - \bar{\mathbf{J}}_{n_0})\mathbf{X}_0];$$

$$q_2 = Tr[\mathbf{A}^T \mathbf{X}_0^T(\mathbf{I}_{n_0} - \bar{\mathbf{J}}_{n_0})\mathbf{X}_0 \mathbf{A}] + Tr[\mathbf{X}^T \mathbf{A}^T \mathbf{X}_0^T(\mathbf{I}_{n_0} - \bar{\mathbf{J}}_{n_0})\mathbf{X}_0 \mathbf{A} \mathbf{X}].$$

Here \mathbf{X} and \mathbf{X}_0 are design matrices for the training and test sets, respectively, and $\mathbf{A} = \mathbf{X}^T(\mathbf{X}\mathbf{X}^T + \lambda\mathbf{I})^{-1}$. The r-score together with M-PGV, EI-PGV and EI-PGV-fwd will be compared to determine the training data to update the prediction model in our strategy.

The Assessment indices

We use the estimates of $CSI_{(j)}$ from the whole phenotype and genotype data as the true $CSI_{(j)}$ values in this study.

CIP@ k

Suppose that the breeder hopes to identify the top k individuals for the true $CSI_{(j)}$ values. Let T_s be the set consisting of the top k individuals for the estimated $CSI_{(j)}$ values. Also, let k_s be the number of individuals which are exactly among the top k individuals for the true $CSI_{(j)}$ values. Then, correctly identified proportion (CIP@ k) is defined as

$$CIP@k = \frac{k_s}{k}.$$

NDCG@ k

Blondel et al. (2015) promoted the use of NDCG (normalized discounted cumulative gain) to measure the ability of various genomic selection strategies to select the top k individuals for the true $CSI_{(j)}$ values. The NDCG has been commonly used to measure the ability of search engines to retrieve highly relevant documents in the top search result (Jarelin and Kekalainen, 2000).

Let $CSI_{(1)} \geq CSI_{(2)} \geq \dots \geq CSI_{(n)}$ be the true $CSI_{(j)}$ values sorted in decreasing order, where $\pi = (\pi_1, \pi_2, \dots, \pi_n)$ is a permutation of $\pi_0 = (1, 2, \dots, n)$. Also, let $\hat{\mathbf{h}}_0 = (\widehat{CSI}_{(1)}, \widehat{CSI}_{(2)}, \dots, \widehat{CSI}_{(n)})$ be the estimated vector of $\mathbf{h}_0 = (CSI_{(1)}, CSI_{(2)}, \dots, CSI_{(n)})$. Then, the DCG score at position k of the predicted ranking is defined as

$$DCG@k(\mathbf{h}_0, \pi(\hat{\mathbf{h}}_0)) = \sum_{j=1}^k f(CSI_{(\pi_j)})d(j)$$

and the DCG score at position k of the ideal ranking is defined as

$$DCG@k(\mathbf{h}_0, \pi_0(\mathbf{h}_0)) = \sum_{j=1}^k f(CSI_{(j)})d(j)$$

where $f(CSI_{(j)})$ is a monotonically increasing gain function and $d(j)$ is a monotonically decreasing discounted function. We consider that $f(CSI_{(j)}) = CSI_{(j)}$ (linear gain) and

$$d(j) = \frac{1}{\log_2(j+1)}.$$

The NDCG score at position k for the selection strategy is then defined as

$$NDCG@k(\mathbf{h}_0, \hat{\mathbf{h}}_0) = \frac{DCG@k(\mathbf{h}_0, \pi(\hat{\mathbf{h}}_0))}{DCG@k(\mathbf{h}_0, \pi_0(\mathbf{h}_0))}.$$

The NDCG score ranges between 0 to 1.



Iterative strategy

Step 0: Select n_0 individuals as an initial training set according to the r-score method, denoted by \mathbf{S}_0 . Initialize $n_{tr} \leftarrow n_0$ and $\mathbf{S}_{tr} \leftarrow \mathbf{S}_0$, where \mathbf{S}_{tr} denotes the current training set and n_{tr} is its sample size.

Step 1: Standardize the phenotypic data of training set, then perform the standardized multiple-trait GBLUP model, and yield the estimated values $\hat{\boldsymbol{\mu}}_c$, $\hat{\boldsymbol{\Sigma}}_g$, $\hat{\boldsymbol{\Sigma}}_e$ and $\hat{\boldsymbol{g}}_{c1}$.

Step 2: Estimate predicted genotypic value of the non-phenotyped set, denoted $\hat{\boldsymbol{g}}_{c2}$, and calculate $\widehat{CSI}_{(j)}$ by $\hat{g}_{ci(j)}$ for $i = 1, 2, \dots, t$; $j = 1, 2, \dots, n_i$.

Step 3: Calculate $CIP@k$ and $NDCG@k$ for a top set of all individuals according to the $\widehat{CSI}_{(j)}$. These two indices are used to evaluate the accuracy for the detection of the superior individuals.

Step 4: Select n_{sel} additional training set individuals from the non-phenotyped set,

denoted \mathcal{S}_{sel} , according to four different acquisition functions, r-score, M-PGV, EI-PGV and EI-PGV-fwd. And add those new training set to the current training set. That is, the union of \mathcal{S}_{tr} and \mathcal{S}_{sel} makes the new training set, expressed as $\mathcal{S}_{tr} \leftarrow \mathcal{S}_{tr} \cup \mathcal{S}_{sel}$. Similarly, $n_{tr} \leftarrow n_{tr} + n_{sel}$. Go to step 1.

Criteria Comparison Based on Real Datasets

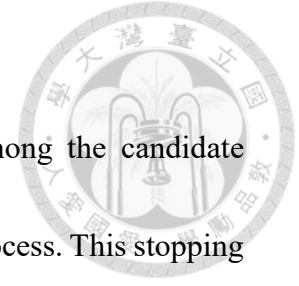
There are three datasets used to demonstrate the iterative strategy and to compare the selection criteria by assessment index based on true $CSI_{(j)}$ values defined above.

The size of starting training set n_0 and training set selected at each batch n_{sel} were the same. Here, we considered the following five cases: (i) $n_0 = n_{sel} = 30$ for the 44k rice dataset, denoted 44k_30; (ii) $n_0 = n_{sel} = 10$ for the tropical rice breeding lines dataset, denoted TR_10; (iii) $n_0 = n_{sel} = 30$ for the tropical rice breeding lines dataset, denoted TR_30; (iv) $n_0 = n_{sel} = 50$ for the tropical rice breeding lines dataset, denoted TR_50; (v) $n_0 = n_{sel} = 30$ for the wheat dataset, denoted wheat_30.

Also, we analyzed different scenarios setting for each dataset as shown in Table 1. There are three scenarios for 44k rice dataset, four scenarios for tropical rice breeding lines dataset and one scenario for wheat dataset. The procedures of Steps 0 to 4 were repeated 30 times for each case and each scenario.

The Stopping Rule for the Iterative Strategy

For a real dataset, the true $CSI_{(j)}$ values are unknown among the candidate population, so we need a stopping rule for the iterative searching process. This stopping rule is according to the EI values in each batch. If the box-plot for a batch approaches 0, then the searching process can stopping. In other word, there is no more improvement made by adding phenotyped individuals to update GBLUP model, whereas the EI values gets equal to 0.



Results



Criteria Comparison Based on Assessment Indices

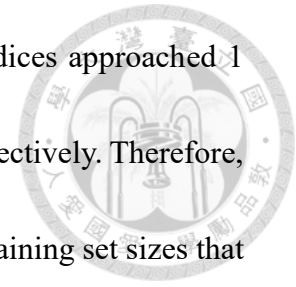
The 44k rice dataset

The results of case (i) for the 44k rice dataset were displayed in Figure 1. In the figure, M-PGV criterion had a better performance than the other criteria before the 5th batch. Around the 5th batch, EI-PGV and EI-PGV-fwd criteria would outperform the other criteria and approached 1 on the $NDCG@k$, regardless of whether the number of selected individuals k is equal to 1, 5 or 10. And $CIP@10$ values are greater than 0.9 around the 6th batch.

The tropical rice dataset

Cases (ii), (iii) and (iv) were the analyses of the tropical rice breeding lines dataset with different batch size setting. The results of case (ii) were displayed in Figure 2. In the figure, M-PGV and r-score criteria had a better performance in the early batches for all the scenarios, regardless of that we used $CIP@k$ or $NDCG@k$ to display. Around the 8th batch, EI-PGV and EI-PGV-fwd criteria would outperform the other two criteria. In the figure, the $NDCG$ indices would be approached 1 around the 16th batch, that's a position which could stop the strategy. For the results of cases (iii) and (iv) were displayed in Figure 3 and 4, we could find the performance of four acquisition functions is similar to the case (ii), and the EI criteria would outperform the other methods after

the training set size reached a certain number. And the NDCG indices approached 1 around the 5th batch and the 3rd batch for case (ii) and case (iii), respectively. Therefore, whether the setting of the size of batch with the same dataset, the training set sizes that made model have enough estimation ability are almost the same.



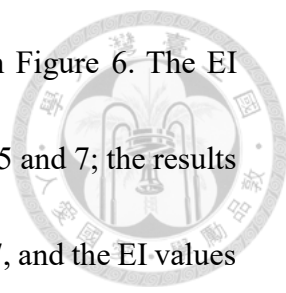
The wheat dataset

The results of case (v) for the wheat dataset were displayed in Figure 5. Here we used this dataset to demonstrate the average of the four responses. The NDCG indices approached 1 around the 13th batch and the 15th batch with EI-PGV and EI-PGV-fwd criteria. The EI methods have better performance than other criteria.

To compare the results of case (i), (iii) and (v), in which the size of initial training set and the batch size are both fixed at 30. For the 44k rice and tropical rice datasets, they used less than half of the size of population to make the model estimate well; but the wheat dataset used more than half of the size of population to achieve the same performance as the other datasets.

The Stopping Rule for the Iterative Strategy

To evaluate the stopping rule for the iterative strategy, we observed the EI values for EI-PGV-fwd, EI-PGV and M-PGV criteria in boxplot, and find the batches whose values approached 0.



The results of case (i) for the 44k rice dataset were shown in Figure 6. The EI values for the criteria approached 0 approximately between batches 5 and 7; the results of case (ii) for the tropical rice breeding lines were shown in Figure 7, and the EI values approached 0 approximately between batches 10 and 12; the results of case (iii) for the tropical rice breeding lines were shown in Figure 8, and the EI values approached 0 approximately between batches 5 and 7; the results of case (iv) for the tropical rice breeding lines were shown in Figure 9, and the EI values approached 0 approximately between batches 3 and 4; the results of case (v) for the wheat dataset were shown in Figure 10, and the EI values approached 0 approximately between batches 14 and 16. These results above are consistent with the results of assessment indices, that the batches were chosen to stop the strategy.

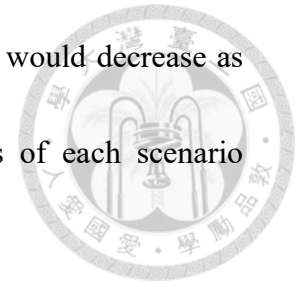
The True Genotypic Values for Specific Batches

To illustrate that the top 10 individuals were selected for specific batches, we transformed the true $CSI_{(j)}$ values back to the true genotypic values, and evaluated the average values of 30 repetitions for each trait. The specific batches were determined by the boxplot of stopping rule for each case.

The 44k rice dataset

The result for case (i) is shown in Table 2. The two traits FTAA and PLHE follow

“the smaller the better”, so we expected the true genotypic values would decrease as the weight increases. From the table, the true genotypic values of each scenario completely follow the weights set in Table 1(a).



The tropical rice dataset

For cases (ii), (iii) and (iv), the trait YLD follows “the larger the better”, and the traits PH and FT follow “the smaller the better”. From Table 3 for case (ii), the results of EI-PGV-fwd and EI-PGV criteria follow the weights set in Table 2(b), although there is some slight deviation on trait FT for some batches, the difference in the values is very small among the scenarios; the result of M-PGV criterion completely follow the weights set in Table 2(b) until the 12th batch on trait YLD. And cases (iii) and (iv) have the same results with the results of case (ii) on the EI-PGV-fwd and EI-PGV criteria, which are displayed in Table 4 and 5, respectively; the result of M-PGV criterion has the same performance with the other criteria in cases (iii) and (iv).

The wheat dataset

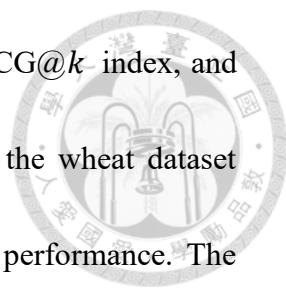
The weights of case (v) we set here represent the highest average yield in all environments. However, we just used this dataset to demonstrate the average of the four responses in this study, and the result of case (v) was displayed in Table 6. GY_E1 doesn't follow the weights in the performance of different batches and acquisition functions, but the performance of other responses follows the setting well.

Discussion



In this study, we applied the r-score method instead of random sampling used in Tanaka and Iwata (2018) and Shen and Liao (2019). The initial training set determined by random sampling doesn't consider any information from dataset. On the other hand, the r-score method considers the information of genotypes, and determines an optimal training set for improving model estimation ability. We re-performed the strategy for cases (i), (iii) and (v) with using the random sampling to determine the initial training set, then calculated the average of CIP@10 and NDCG@10 at the first batch with the 30 repetitions, the results are presented in Table 7. The results show that the r-score can indeed make the model have better estimation ability at the first batch, and r-score is even much better than random sampling under Tropical rice dataset. Using r-score at the first batch can make the strategy more efficient, in other words, it has better model prediction ability and it can select superior genotypes in earlier batches.

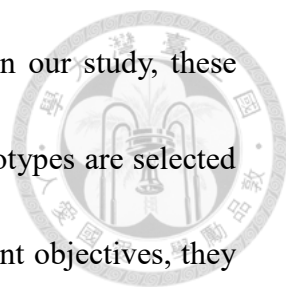
The assessment indices, CIP@ k and NDCG@ k , seem to be similar, but NDCG not only compares the top k individuals like CIP, but also considers their ranking and the individuals other than the top k . We find the EI-PGV and EI-PGV-fwd criteria seem to be more stable on training set determination for model estimation, even they didn't have the best performance in earlier batches, but they could approach 1 faster than other criteria by NDCG@ k . For the 44k rice and Tropical rice datasets, the training set size



just requires less than half of the dataset size according to the $NDCG@k$ index, and it's enough to make the model have great prediction ability; but the wheat dataset requires more than half of the dataset size to achieve the same performance. The possible reason for the wheat dataset is that the number of markers is too small, resulting in inaccurate model estimation. The 44k rice dataset needs about 150 out of 373 training individuals for performing the strategy according to Figure 6; the tropical rice dataset needs about 150 to 160 out of 328 training individuals for performing the strategy according to Figures 7, 8, and 9; but the wheat dataset needs about 420 out of 599 training individuals for performing the strategy according to Figure 10.

We used a stopping rule for the strategy if the true genotypic values are unknown. Comparison of plot for the assessment indices and boxplot for stopping rule, their results are consistent. It means that this stopping rule can really help us decide when to stop the strategy. However, we need to further study the feasibility of CSI. The resulting tables of the true genotypic values show that the values indeed follow the scenarios in this study. In general, EI-PGV and EI-PGV-fwd criteria have better performance in the sequential phenotyping strategy, it can select superior genotypes more effectively and efficiently than other criteria. And CSI can indeed be effectively applied to the multiple-trait selection.

The EI-PGV and EI-PGV-fwd criteria performed in Tanaka and Iwata (2018) and



Shen and Liao (2019) were used to explore the best individual. In our study, these criteria are used to determine the training set, and the superior genotypes are selected by $\widehat{CSI}_{(j)}$. However, even if the EI criteria were applied for different objectives, they were shown to have a great performance on the strategy of this study.

In general, this strategy really helps us to detect superior genotypes in an efficient way. We used several acquisition functions and two assessment indices to find best training set determination methods for this strategy. Finally, we observed that EI criteria have the most stable and better performance than other criteria. From this study, it showed that EI methods not only have a great performance in exploring the best genotype (Shen and Liao, 2019), but also have a good performance in selecting a good training population for model estimation. In our experience, the computing time for performing our proposed iterative process could mainly depend on the number of traits and the number of individuals in the candidate population. We compare the computing time for Cases (i), (iii), and (v) which all have the batch size of 30. It required about 4 to 5 hours for completing the 44k rice dataset analysis with 2 traits and 373 individuals in Case (i). Similarly, it required about 1 day, and 6 to 7 days for the tropical rice dataset with 3 traits and 328 individuals in Case (iii), and the wheat dataset with 4 traits and 599 individuals in Case (v), respectively. Developing an effective algorithm to reduce the computing time will be our future study.

References



Blondel, M., A. Onogi, H. Iwata, and N. Ueda, (2015). A Ranking Approach to Genomic Selection. *PLOS One* 10, e0128570.

Covarrubias-Pazaran, G. (2016). Genome-Assisted Prediction of Quantitative Traits Using the R Package sommer. *PLOS One* 11, e0156744.

Crossa, J., G. de los. Campos, P. Pérez, D. Gianola, J. Burgueño, et al. (2010). Prediction of Genetic Values of Quantitative Traits in Plant Breeding Using Pedigree and Molecular Markers. *Genetics* 186, 713–724.

Gaynor, R.C., G. Gorjanc, A.R. Bentley, E.S. Ober, P. Howell, et al. (2017). A two-part strategy for using genomic selection to develop inbred lines. *Crop Sci.* 57: 2372–2386.

Gong, C., J. Peng, and Q. Liu, (2019). Quantile Stein variational gradient descent for batch Bayesian optimization. *Proceedings of the 36th International Conference on Machine Learning, PMLR 97*, 2347-2356.

Guo, G., F. Zhao, Y. Wang, Y. Zhang, L. Du, et al. (2014). Comparison of single-trait and multiple-trait genomic prediction models. *BMC Genetics* 15, 30.

Hayashi, T. and H. Iwata, (2013). A Bayesian method and its variational approximation for prediction of genomic breeding values in multiple traits. *BMC Bioinformatics* 14, 34.



Hyten, D. L., Q. Song, Y. Zhu, I.Y. Choi, R.L. Nelson, et al. (2006). Impacts of genetic bottlenecks on soybean genome diversity. *Proceedings of the National Academy of Sciences* 103, 16666–16671.

Jia, Y. and J.L. Jannink, (2012). Multiple-Trait Genomic Selection Methods Increase Genetic Value Prediction Accuracy. *Genetics* 192, 1513–1522.

Khoury, C. K., A.D. Bjorkman, H. Dempewolf, J. Ramirez-Villegas, L. Guarino, et al. (2014). Increasing homogeneity in global food supplies and the implications for food security. *Proceedings of the National Academy of Sciences* 111, 4001–4006.

McCouch, S., G.J. Baute, J. Bradeen, P. Bramel, P.K. Bretting, et al. (2013). Feeding the future. *Nature* 499, 23–24.

Meuwissen, T.H.E., B.J. Hayes, and M.E. Goddard, (2001). Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. *Genetics* 157, 1819.

Ou, J.-H. and C.T. Liao, (2019). Training set determination for genomic selection. *Theor. Appl. Genet.* 132, 2781–2792.

Reif, J.C., P. Zhang, S. Dreisigacker, M.L. Warburton, M. van Ginkel, et al. (2005). Wheat genetic diversity trends during domestication and breeding. *Theor. Appl. Genet.* 110, 859–864.

Schulthess, A.W., Y. Wang, T. Miedaner, P. Wilde, J.C. Reif, et al. (2016). Multiple-trait- and selection indices-genomic predictions for grain yield and protein content

in rye for feeding purposes. *Theor. Appl. Genet.* 129, 273–287.

Searle, S.R. (1982). *Matrix algebra useful for statistics*. New York: JOHN WILEY & SONS.



Shahriari, B., K. Swersky, Z. Wang, R.P. Adams, N. de Freitas, (2016). Taking the Human Out of the Loop: A Review of Bayesian Optimization. *Proc. IEEE* 104, 148–175.

Shen. (2019). Identification of the best genotype from a large candidate set. Degree Thesis of National Taiwan University.

Spindel, J., H. Begum, D. Akdemir, P. Virk, B. Collard, et al. (2015). Genomic Selection and Association Mapping in Rice (*Oryza sativa*): Effect of Trait Genetic Architecture, Training Population Composition, Marker Number and Statistical Model on Accuracy of Rice Genomic Selection in Elite, Tropical Rice Breeding Lines. *PLOS Genetics* 11, e1004982.

Tanaka, R. and H. Iwata, (2018). Bayesian optimization for genomic selection: a method for discovering the best genotype among a large number of candidates. *Theor. Appl. Genet.* 131, 93–105.

Tanksley, S.D. and S.R. McCouch, (1997). Seed Banks and Molecular Maps: Unlocking Genetic Potential from the Wild. *Science* 277, 1063–1066.

Tester, M. and P. Langridge, (2010). Breeding Technologies to Increase Crop

Production in a Changing World. *Science* 327, 818–822.



VanRaden, P.M. (2008). Efficient Methods to Compute Genomic Predictions. *Journal of Dairy Science* 91, 4414–4423.

Zhang, A., H. Wang, Y. Beyene, K. Semagn, Y. Liu, et al. (2017). Effect of Trait Heritability, Training Population Size and Marker Density on Genomic Prediction Accuracy Estimation in 22 bi-parental Tropical Maize Populations. *Front. Plant Sci.* 8, 1916.

Zhao, K., C.W. Tung, G.C. Eizenga, M.H. Wright, M.L. Ali, et al. (2011). Genome-wide association mapping reveals a rich genetic architecture of complex traits in *Oryza sativa*. *Nat. Commun.* 2, 467.

Table 1. The specified weight for trait i of each dataset.

(a) 44 rice dataset

Scenarios	Weight (p_i)	
	FTAA	PLHE
1	1	0
2	0	1
3	0.5	0.5

(b) Tropical rice breeding lines dataset

Scenarios	Weight (p_i)		
	YLD	PH	FT
1	1	0	0
2	0.7	0.3	0
3	0.7	0	0.3
4	0.7	0.15	0.15

(c) Wheat dataset

Scenarios	Weight (p_i)			
	GD_E1	GD_E2	GD_E3	GD_E4
1	0.25	0.25	0.25	0.25



Table 2. For the 44k rice dataset with the batch size is set to 30, the average of true genotypic values of the individuals which are selected on specific batches according to the stopping rule over the 30 repetitions.

44k_30 scenario	True Genotypic Value					
	EI-PGV-fwd		EI-PGV		M-PGV	
(Batch 4)	FTAA	PLHE	FTAA	PLHE	FTAA	PLHE
1	64.29	97.97	64.30	98.25	64.45	97.91
2	86.26	87.73	85.18	87.88	78.62	87.03
3	67.10	92.23	67.36	92.06	67.14	93.18
(Batch 5)	FTAA	PLHE	FTAA	PLHE	FTAA	PLHE
1	64.07	97.95	64.04	98.01	64.42	97.97
2	82.68	86.65	83.17	86.88	78.52	86.72
3	66.43	92.53	66.69	92.23	66.99	92.36
(Batch 6)	FTAA	PLHE	FTAA	PLHE	FTAA	PLHE
1	64.02	98.01	64.01	98.01	64.10	98.03
2	79.86	86.23	79.41	86.22	78.45	86.44
3	66.35	92.59	66.40	92.49	67.34	91.34
(Batch 7)	FTAA	PLHE	FTAA	PLHE	FTAA	PLHE
1	64.01	98.04	64.01	98.01	64.10	97.89
2	78.18	86.04	78.11	86.07	77.94	86.22
3	66.25	92.72	66.32	92.61	67.11	91.48
(Batch 8)	FTAA	PLHE	FTAA	PLHE	FTAA	PLHE
1	64.01	98.03	64.01	98.01	64.01	98.03
2	77.68	86.08	77.61	86.07	77.23	86.18
3	66.25	92.71	66.25	92.70	67.07	91.53

FTAA: flowering time at Arkansas; PLHE: plant height; M-PGV: the strategy based on the mean of predicted genotypic values; EI-PGV: the strategy with the expected improvement criterion based on distribution of predicted genotypic values; EI-PGV-fwd: the strategy with the forward expected improvement criterion based on distribution of predicted genotypic values.

Table 3. For the tropical rice breeding lines dataset with the batch size is set to 10, the average of true genotypic values of the individuals which are selected on specific batches according to the stopping rule over the 30 repetitions.

TR_10	True Genotypic Value								
scenario	EI-PGV-fwd			EI-PGV			M-PGV		
(Batch 9)	YLD	PH	FT	YLD	PH	FT	YLD	PH	FT
1	5464.35	107.36	85.60	5470.46	107.08	85.88	5421.93	109.14	86.80
2	5454.64	105.95	86.04	5429.91	105.45	86.16	5371.41	103.52	85.81
3	5435.44	109.76	85.04	5435.94	108.36	85.06	5430.98	106.20	85.01
4	5462.45	107.04	85.37	5464.21	106.85	85.80	5399.88	104.75	85.12
(Batch 10)	YLD	PH	FT	YLD	PH	FT	YLD	PH	FT
1	5481.60	107.66	85.73	5480.57	107.39	85.95	5428.80	109.30	86.91
2	5459.02	105.46	85.93	5465.49	105.65	85.95	5375.17	103.48	85.77
3	5450.73	109.99	85.13	5457.62	108.56	85.27	5438.37	106.58	85.09
4	5477.00	107.00	85.45	5474.94	107.01	85.93	5411.35	104.70	85.17
(Batch 11)	YLD	PH	FT	YLD	PH	FT	YLD	PH	FT
1	5485.25	107.38	85.75	5486.14	107.60	85.98	5435.39	109.03	86.80
2	5461.65	105.09	85.73	5467.74	105.45	85.82	5383.78	103.30	85.61
3	5457.02	109.88	85.03	5473.53	108.56	85.30	5445.17	106.70	84.96
4	5481.12	106.90	85.61	5479.31	107.08	85.93	5424.62	104.75	85.01
(Batch 12)	YLD	PH	FT	YLD	PH	FT	YLD	PH	FT
1	5489.30	107.57	85.78	5487.19	107.52	85.93	5450.63	108.88	86.75
2	5460.67	104.82	85.65	5458.60	104.99	85.69	5391.27	103.17	85.55
3	5463.99	109.23	85.14	5477.24	108.21	85.21	5449.49	106.81	84.83
4	5485.62	106.91	85.67	5479.78	106.72	85.85	5433.93	104.88	84.84
(Batch 13)	YLD	PH	FT	YLD	PH	FT	YLD	PH	FT
1	5488.12	107.49	85.79	5487.85	107.76	86.02	5457.92	109.06	86.75
2	5452.54	104.28	85.60	5457.24	104.42	85.57	5401.02	103.01	85.40
3	5474.98	108.48	85.12	5480.55	108.05	85.06	5455.93	106.82	84.82
4	5484.46	106.60	85.50	5481.20	106.19	85.66	5439.40	105.05	84.97

YLD: grain yield; PH: plant height; FT: flowering time; M-PGV: the strategy based on the mean of predicted genotypic values; EI-PGV: the strategy with the expected improvement criterion based on distribution of predicted genotypic values; EI-PGV-fwd: the strategy with the forward expected improvement criterion based on distribution of predicted genotypic values.

Table 4. For the tropical rice breeding lines dataset with the batch size is set to 30, the average of true genotypic values of the individuals which are selected on specific batches according to the stopping rule over the 30 repetitions.

TR_30	True Genotypic Value								
scenario	EI-PGV-fwd			EI-PGV			M-PGV		
(Batch 4)	YLD	PH	FT	YLD	PH	FT	YLD	PH	FT
1	5493.85	107.30	85.77	5499.73	107.43	85.81	5484.73	107.58	86.06
2	5466.46	104.98	85.55	5457.58	104.49	85.54	5426.63	102.39	84.89
3	5466.56	107.85	84.99	5472.43	107.93	84.97	5468.98	106.47	84.74
4	5486.37	106.72	85.55	5479.49	106.37	85.39	5451.39	104.24	84.79
(Batch 5)	YLD	PH	FT	YLD	PH	FT	YLD	PH	FT
1	5503.18	107.68	86.01	5506.64	107.90	86.05	5497.01	107.84	86.19
2	5461.34	104.20	85.43	5452.97	103.66	85.29	5430.89	102.37	84.90
3	5495.06	107.06	84.91	5493.37	107.22	84.82	5474.41	106.84	84.71
4	5488.54	106.01	85.31	5487.99	105.94	85.27	5459.36	104.46	84.74
(Batch 6)	YLD	PH	FT	YLD	PH	FT	YLD	PH	FT
1	5509.42	107.70	86.03	5512.26	107.73	86.05	5505.04	108.36	86.41
2	5446.90	103.02	85.21	5442.35	102.87	85.23	5435.36	102.54	84.94
3	5494.45	106.73	84.57	5497.15	106.86	84.51	5485.12	107.16	84.64
4	5491.08	105.49	85.09	5491.74	105.59	85.02	5458.72	104.25	84.60
(Batch 7)	YLD	PH	FT	YLD	PH	FT	YLD	PH	FT
1	5514.54	107.83	86.06	5518.04	107.89	86.17	5509.58	108.92	86.69
2	5437.02	102.18	84.98	5439.01	102.37	85.06	5434.67	102.28	84.94
3	5496.59	106.73	84.38	5499.20	106.80	84.42	5494.09	107.48	84.61
4	5479.63	104.98	84.55	5484.40	105.09	84.66	5465.23	104.18	84.60
(Batch 8)	YLD	PH	FT	YLD	PH	FT	YLD	PH	FT
1	5520.61	108.11	86.00	5522.22	108.13	86.05	5515.32	108.86	86.55
2	5435.65	102.00	84.91	5437.96	102.07	84.89	5432.88	102.04	84.95
3	5492.05	106.61	84.25	5491.97	106.54	84.26	5497.79	107.32	84.55
4	5454.16	103.44	84.22	5455.13	103.33	84.33	5469.02	104.24	84.57

YLD: grain yield; PH: plant height; FT: flowering time; M-PGV: the strategy based on the mean of predicted genotypic values; EI-PGV: the strategy with the expected improvement criterion based on distribution of predicted genotypic values; EI-PGV-fwd: the strategy with the forward expected improvement criterion based on distribution of predicted genotypic values.

Table 5. For the tropical rice breeding lines dataset with the batch size is set to 50, the average of true genotypic values of the individuals which are selected on specific batches according to the stopping rule over the 30 repetitions.

TR_50	True Genotypic Value								
scenario	EI-PGV-fwd			EI-PGV			M-PGV		
(Batch 2)	YLD	PH	FT	YLD	PH	FT	YLD	PH	FT
1	5462.68	106.66	85.34	5467.87	106.47	85.33	5470.19	107.69	86.12
2	5423.59	105.38	85.93	5420.73	104.93	85.82	5417.25	103.59	85.18
3	5417.13	109.47	85.20	5448.85	108.65	84.97	5442.65	106.95	85.04
4	5434.60	106.87	85.58	5448.67	106.45	85.39	5441.80	104.94	85.02
(Batch 3)	YLD	PH	FT	YLD	PH	FT	YLD	PH	FT
1	5504.23	107.35	85.76	5506.12	107.23	85.71	5494.21	107.99	86.07
2	5459.48	104.41	85.42	5466.83	104.62	85.44	5431.75	102.42	85.01
3	5492.00	107.83	84.79	5499.86	107.68	84.99	5481.58	106.62	84.68
4	5498.24	106.54	85.36	5494.83	106.07	85.22	5463.41	104.36	84.72
(Batch 4)	YLD	PH	FT	YLD	PH	FT	YLD	PH	FT
1	5515.60	107.44	85.76	5516.85	107.45	85.78	5510.10	108.71	86.44
2	5454.11	103.21	85.05	5452.32	103.27	85.21	5433.26	102.16	84.96
3	5501.04	106.90	84.50	5499.44	106.84	84.40	5489.44	107.04	84.47
4	5497.82	105.60	84.94	5494.86	105.49	84.90	5475.62	106.65	84.69
(Batch 5)	YLD	PH	FT	YLD	PH	FT	YLD	PH	FT
1	5523.78	107.96	85.84	5523.22	108.01	85.77	5518.65	108.71	86.53
2	5443.78	102.37	84.87	5444.47	102.38	84.85	5435.10	102.01	84.93
3	5485.93	106.34	84.15	5488.13	106.40	84.19	5494.74	107.10	84.40
4	5465.63	104.21	84.24	5468.12	104.33	84.23	5470.81	104.2	84.45
(Batch 6)	YLD	PH	FT	YLD	PH	FT	YLD	PH	FT
1	5527.90	107.69	85.58	5527.89	107.77	85.71	5526.26	108.20	86.23
2	5434.69	101.91	84.83	5436.31	101.97	84.82	5432.92	101.87	84.90
3	5483.01	106.07	84.01	5483.88	106.11	84.02	5485.68	106.45	84.07
4	5471.40	104.36	84.14	5464.38	103.84	84.18	5469.97	104.21	84.15

YLD: grain yield; PH: plant height; FT: flowering time; M-PGV: the strategy based on the mean of predicted genotypic values; EI-PGV: the strategy with the expected improvement criterion based on distribution of predicted genotypic values; EI-PGV-fwd: the strategy with the forward expected improvement criterion based on distribution of predicted genotypic values.

Table 6. For the wheat dataset with the batch size is set to 30, the average of true genotypic values of the individuals which are selected on specific batches according to the stopping rule over the 30 repetitions.

Wheat_30 scenario	True Genotypic Value											
	EI-PGV-fwd				EI-PGV				M-PGV			
(Batch 12)	GY_E1	GY_E2	GY_E3	GY_E4	GY_E1	GY_E2	GY_E3	GY_E4	GY_E1	GY_E2	GY_E3	GY_E4
1	0.01	0.91	0.93	0.91	0.00	0.90	0.92	0.92	-0.30	0.84	0.88	1.06
(Batch 13)	GY_E1	GY_E2	GY_E3	GY_E4	GY_E1	GY_E2	GY_E3	GY_E4	GY_E1	GY_E2	GY_E3	GY_E4
1	0.01	0.91	0.93	0.92	0.00	0.91	0.93	0.92	-0.28	0.82	0.87	1.07
(Batch 14)	GY_E1	GY_E2	GY_E3	GY_E4	GY_E1	GY_E2	GY_E3	GY_E4	GY_E1	GY_E2	GY_E3	GY_E4
1	-0.01	0.91	0.94	0.94	0.00	0.90	0.93	0.93	-0.25	0.83	0.88	1.06
(Batch 15)	GY_E1	GY_E2	GY_E3	GY_E4	GY_E1	GY_E2	GY_E3	GY_E4	GY_E1	GY_E2	GY_E3	GY_E4
1	-0.03	0.92	0.95	0.96	-0.02	0.91	0.94	0.95	-0.26	0.84	0.88	1.06
(Batch 16)	GY_E1	GY_E2	GY_E3	GY_E4	GY_E1	GY_E2	GY_E3	GY_E4	GY_E1	GY_E2	GY_E3	GY_E4
1	-0.04	0.91	0.95	0.98	-0.02	0.92	0.95	0.96	-0.26	0.86	0.90	1.05

GY_E1: grain yield at environment 1; GY_E2: grain yield at environment 2; GY_E3: grain yield at environment 3; GY_E4: grain yield at environment 4; M-PGV: the strategy based on the mean of predicted genotypic values; EI-PGV: the strategy with the expected improvement criterion based on distribution of predicted genotypic values; EI-PGV-fwd: the strategy with the forward expected improvement criterion based on distribution of predicted genotypic values.

Table 7. Comparison that the average of CIP@10 and NDCG@10 for random sampling and r-score at the first batch with size is fixed at 30.

(a) 44k rice dataset

Scenario	CIP@10		NDCG@10	
	RS	r-score	RS	r-score
1	0.17	0.19	0.39	0.46
2	0.12	0.16	0.69	0.70
3	0.12	0.09	0.55	0.53

(b) Tropical rice breeding lines dataset

Scenario	CIP@10		NDCG@10	
	RS	r-score	RS	r-score
1	0.12	0.30	0.35	0.62
2	0.16	0.42	0.29	0.56
3	0.11	0.26	0.34	0.59
4	0.14	0.35	0.32	0.57

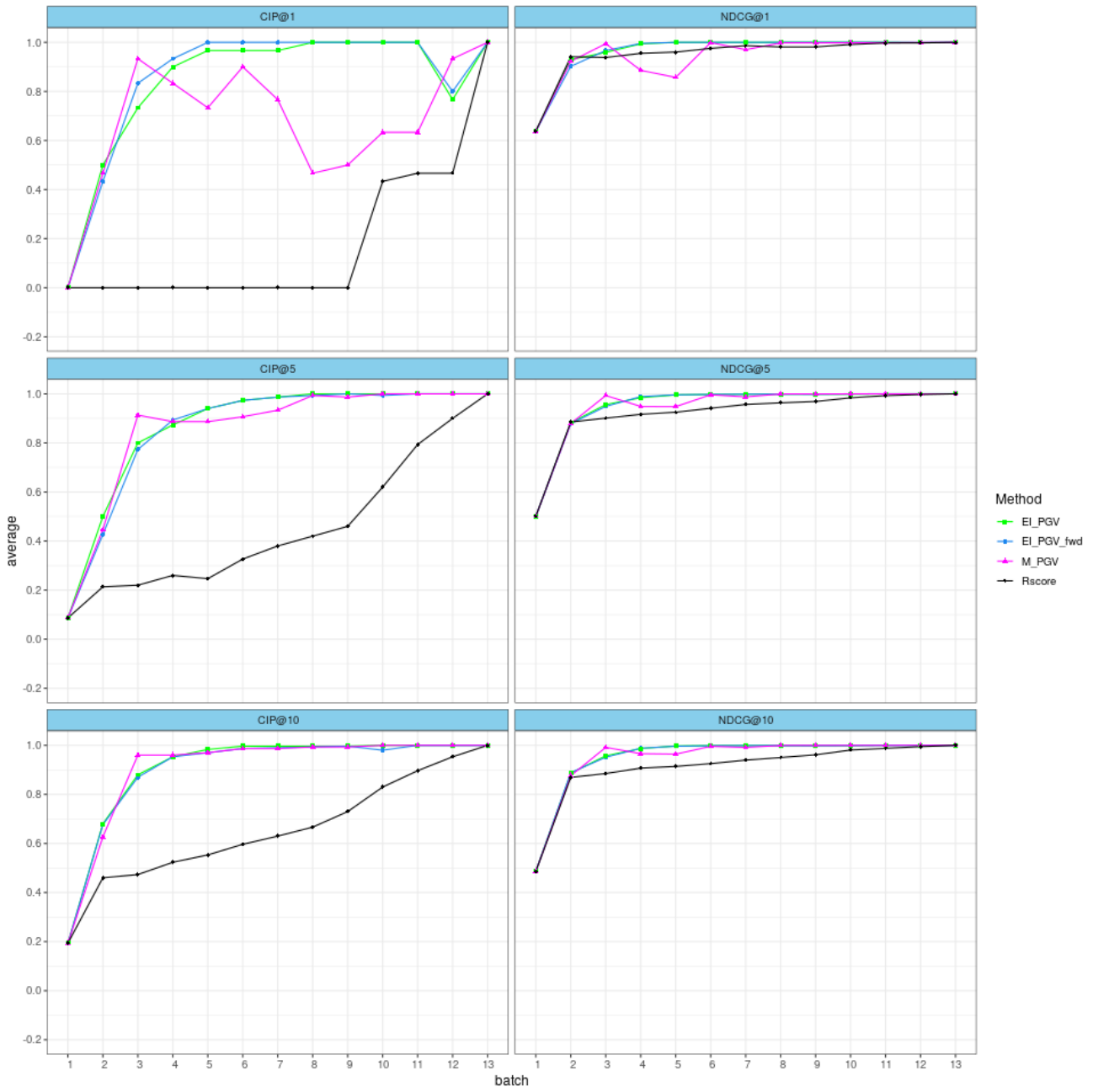
(c) Wheat dataset

scenario	CIP@10		NDCG@10	
	RS	r-score	RS	r-score
1	0.05	0.05	0.30	0.43

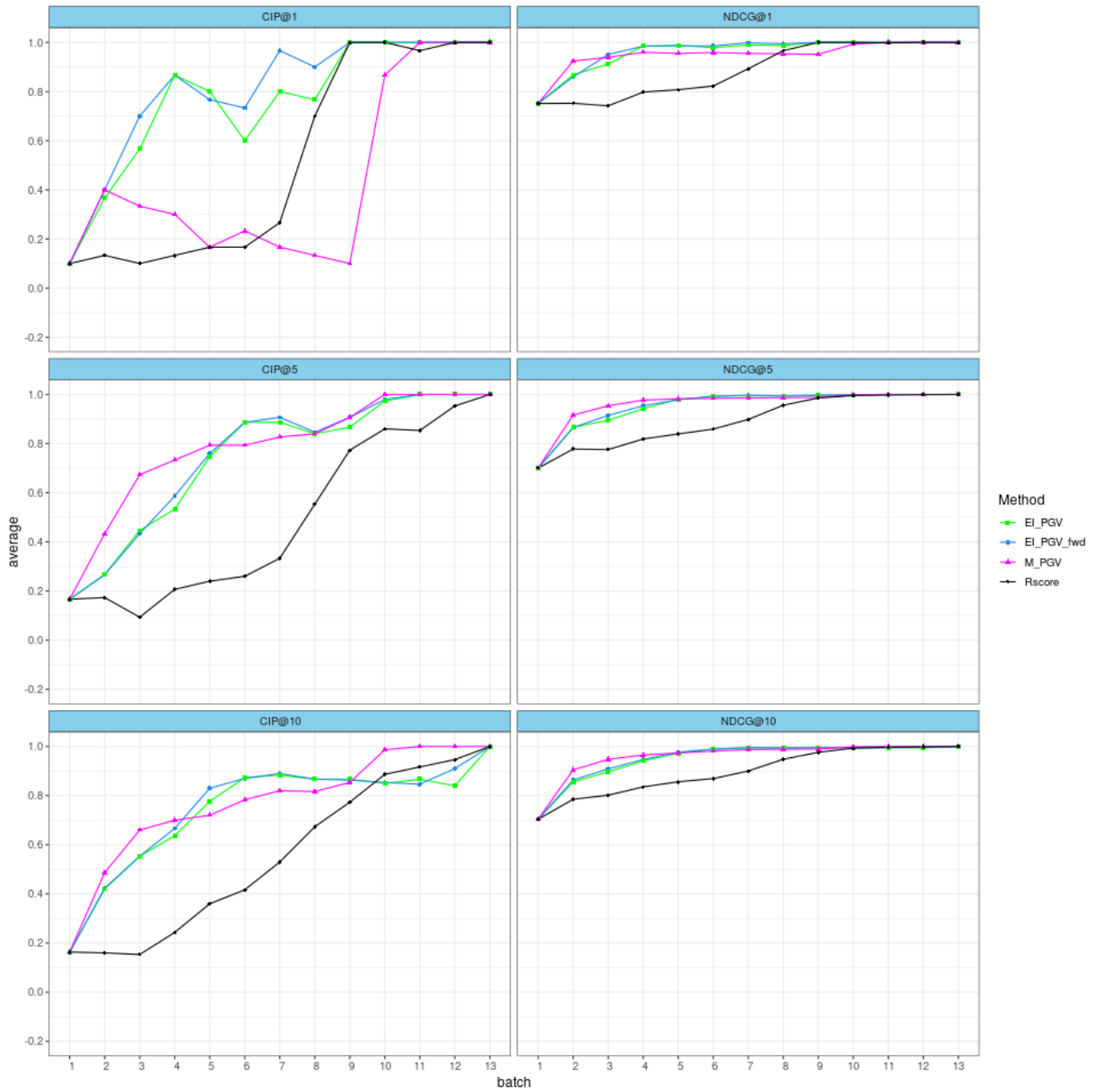
CIP@10: correct identification proportion of top 10 individuals; NDCG@10: normalized discounted cumulative gain of top 10 individuals; RS: random sampling.



(a) Scenario 1



(b) Scenario 2



(c) Scenario 3

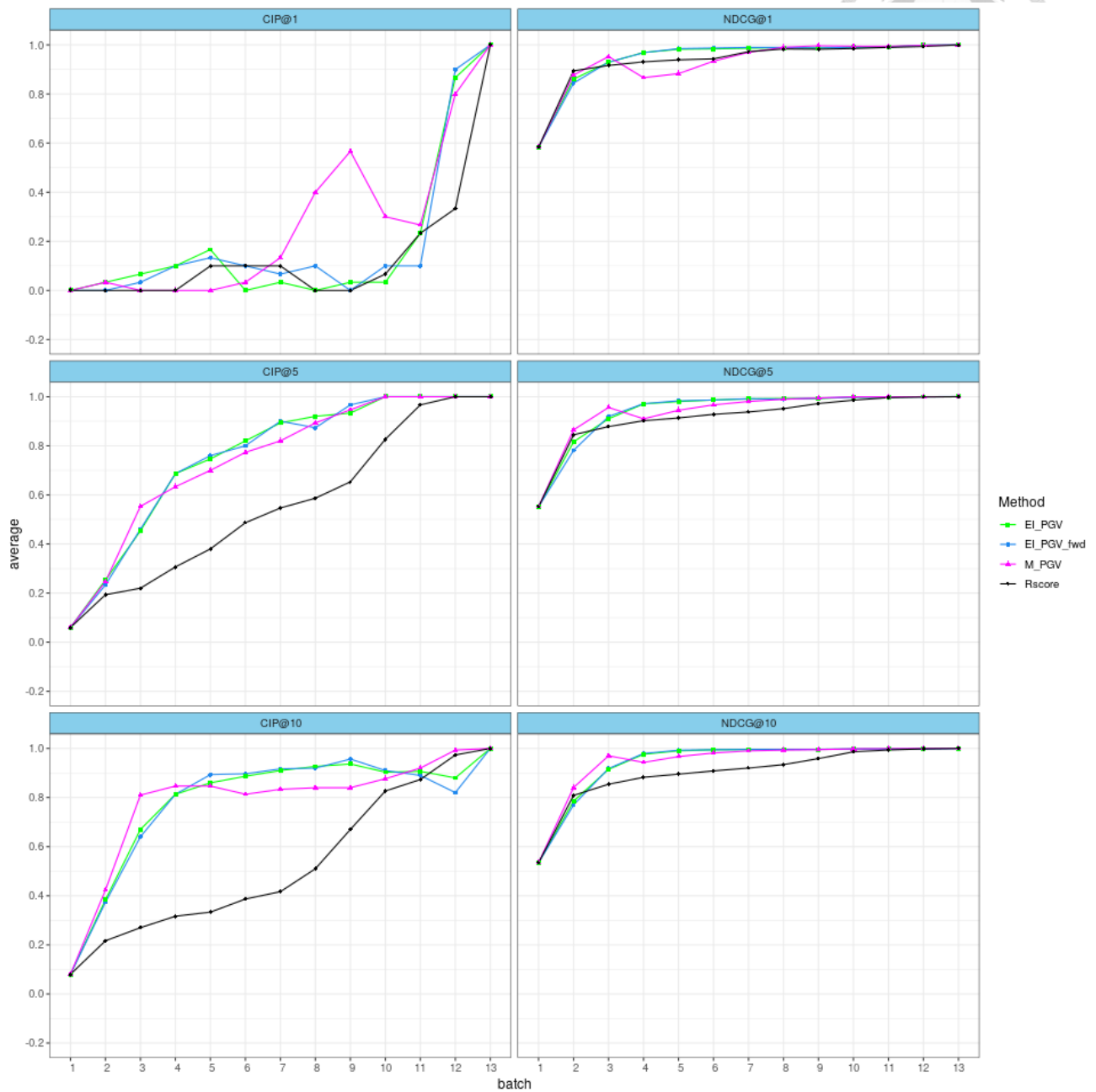
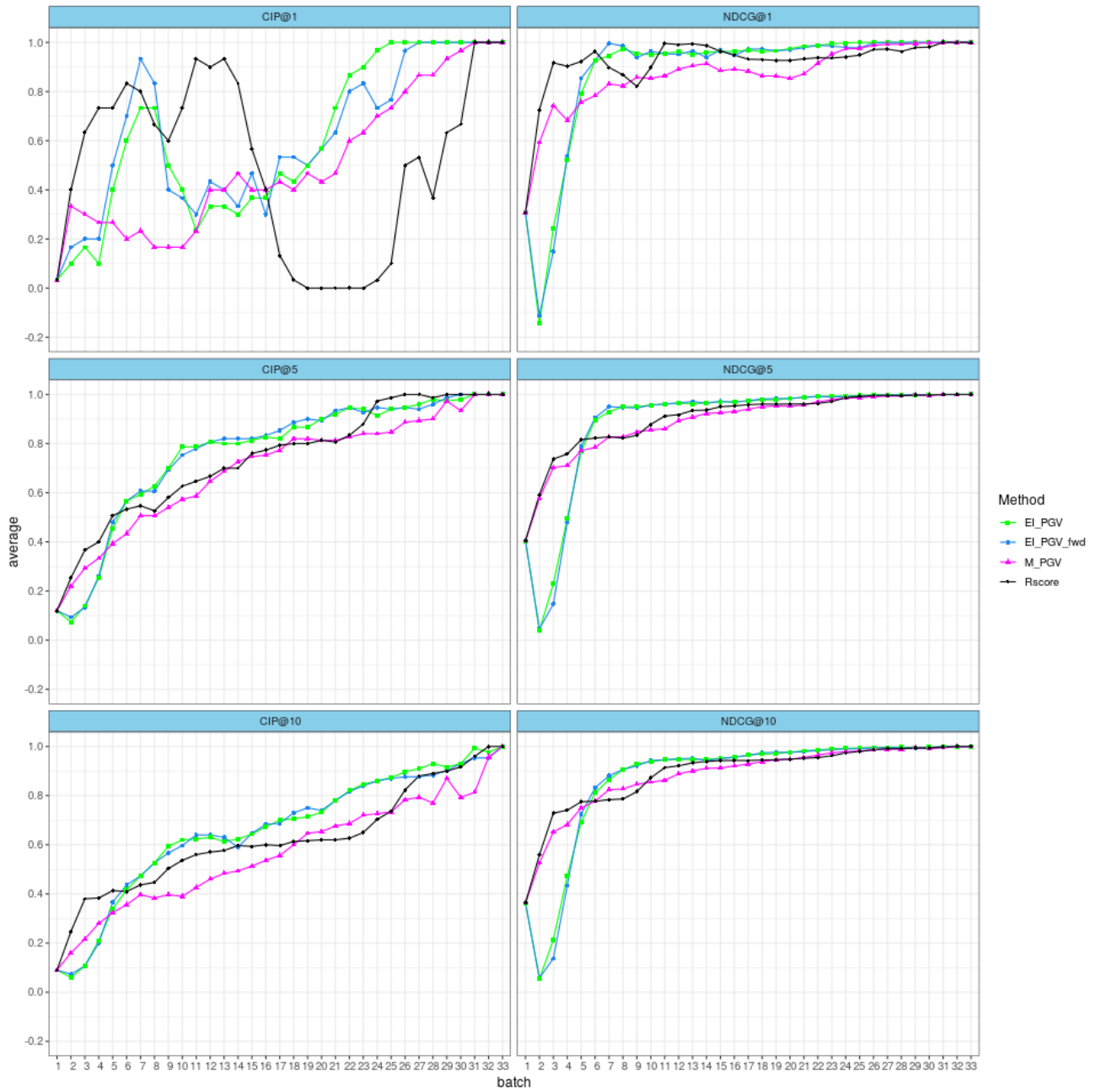
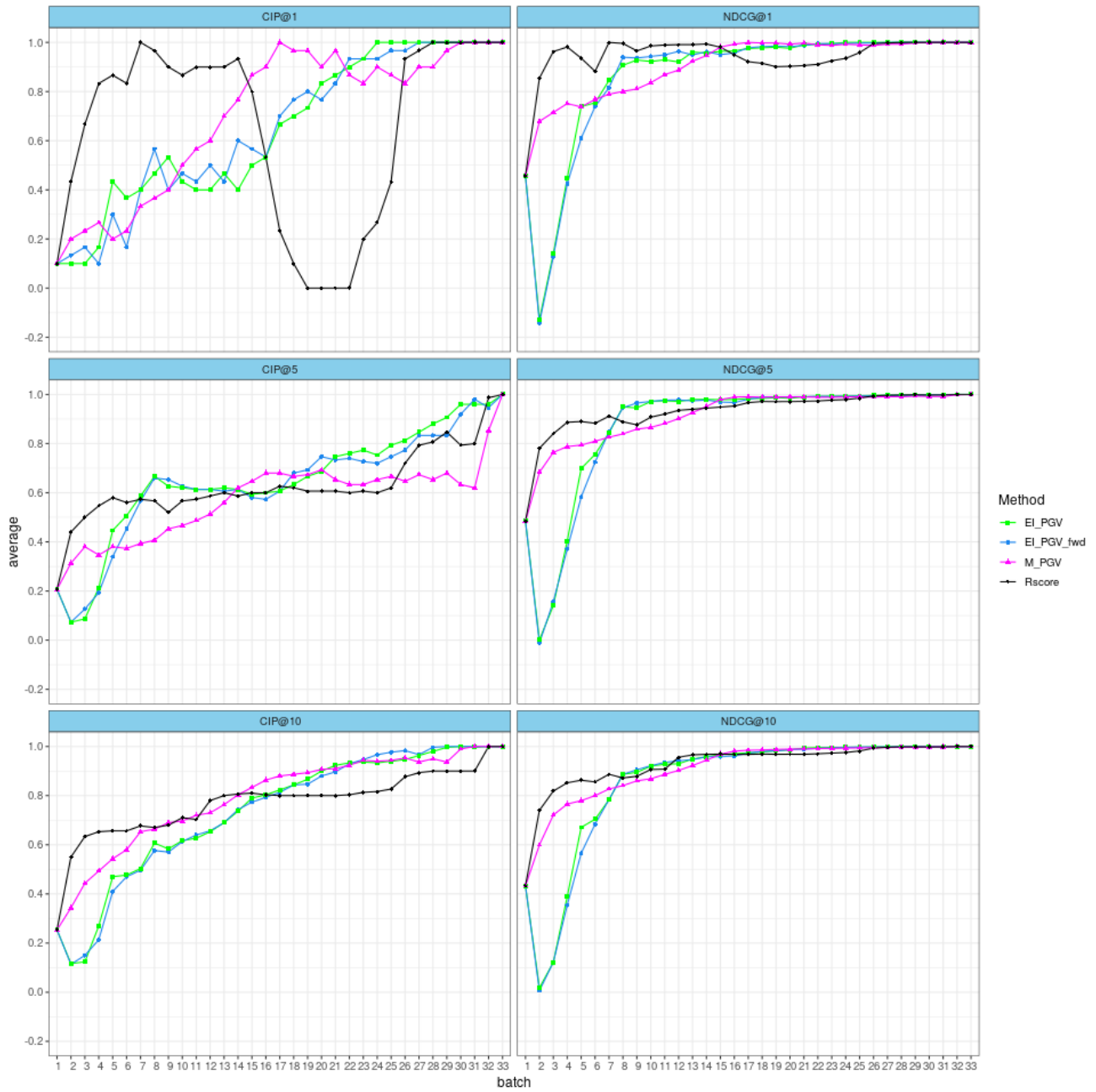


Figure 1. The assessment indices for case (i). For the 44k rice dataset with the batch size is set to 30, the average of correct identification proportion (CIP) and normalized discounted cumulative gain of top k individuals (NDCG@ k), which were calculated at each batch over the 30 repetitions.

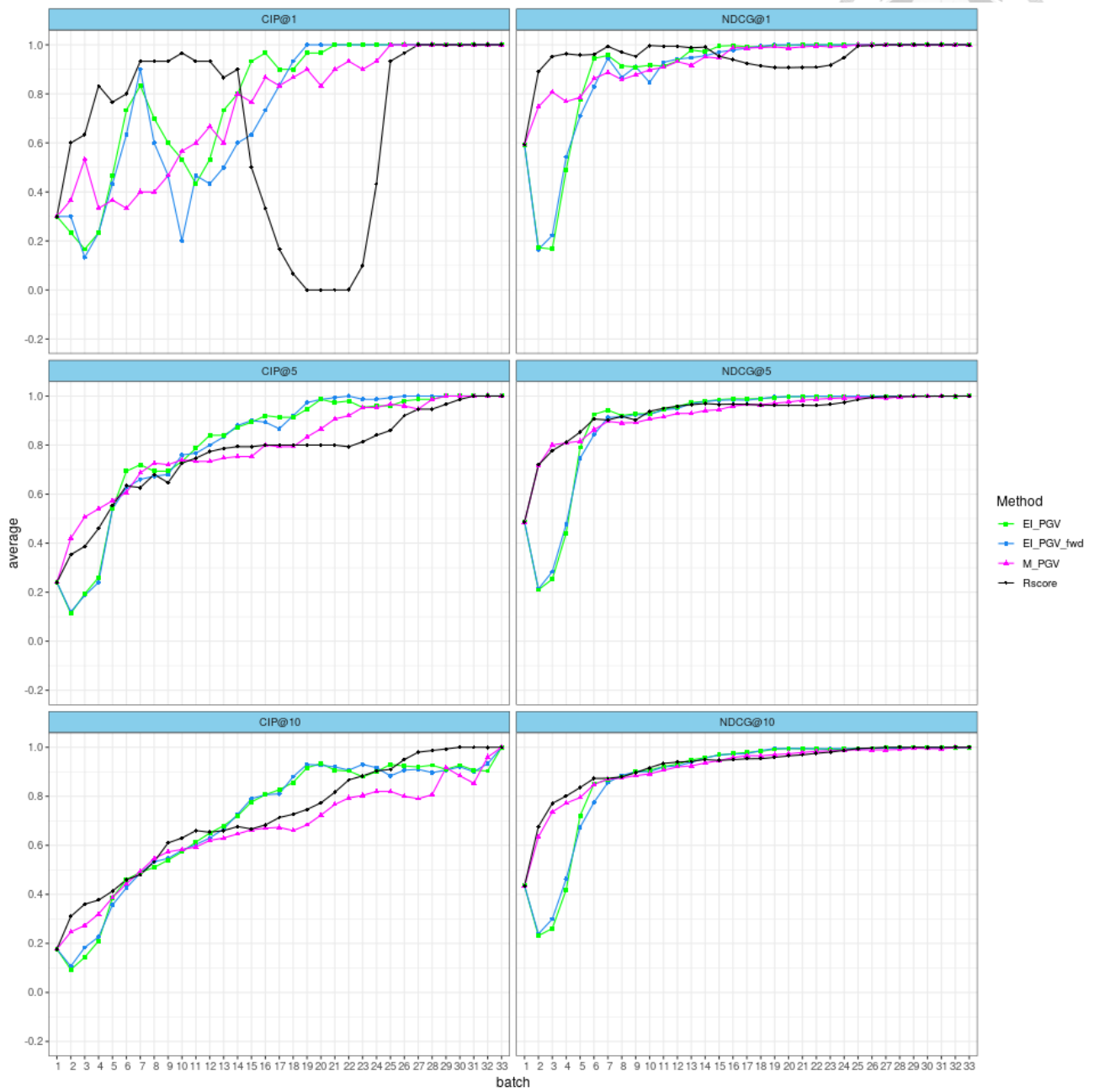
(a) Scenarios 1



(b) Scenario 2



(c) Scenario 3



(d) Scenario 4

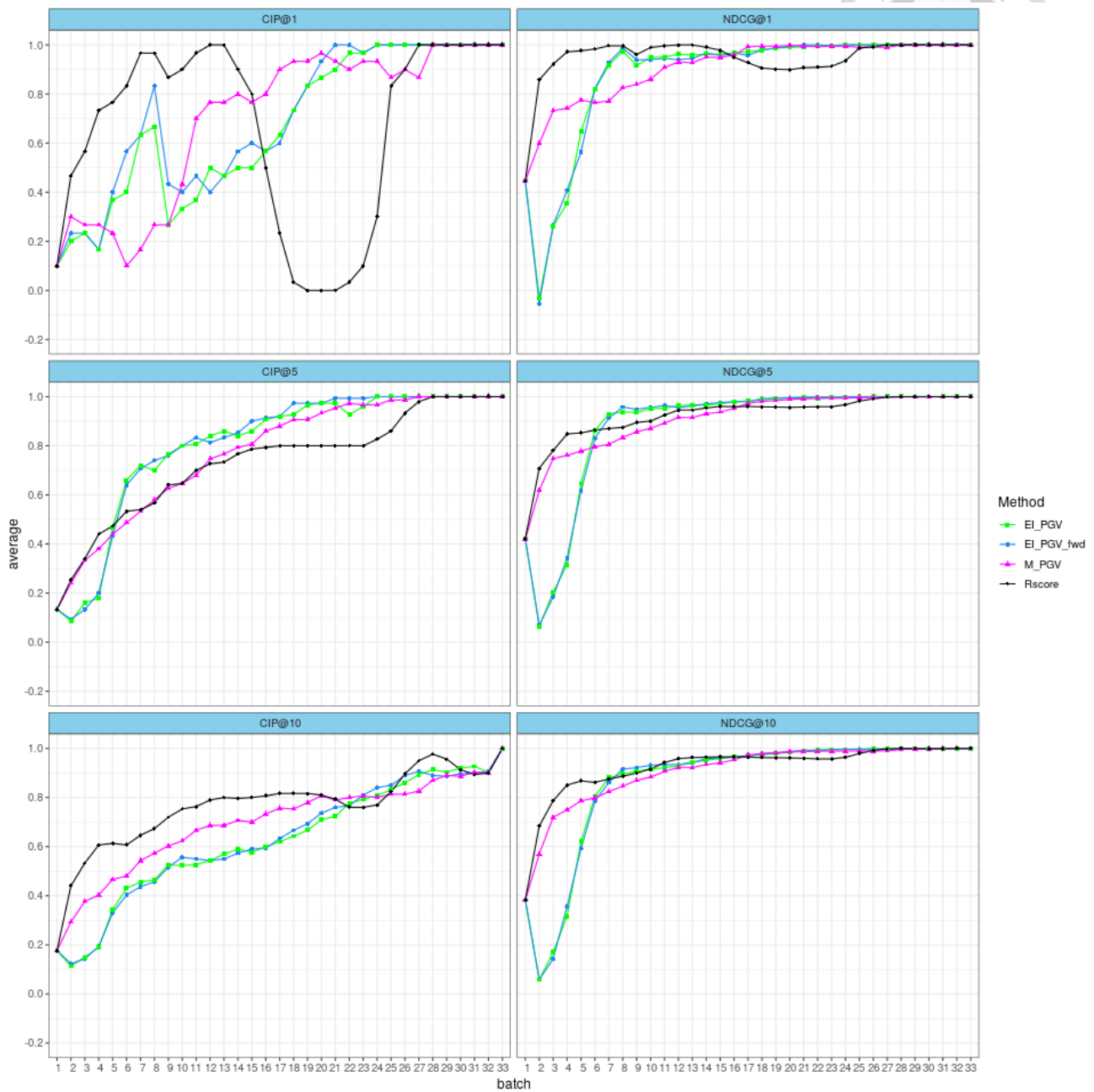
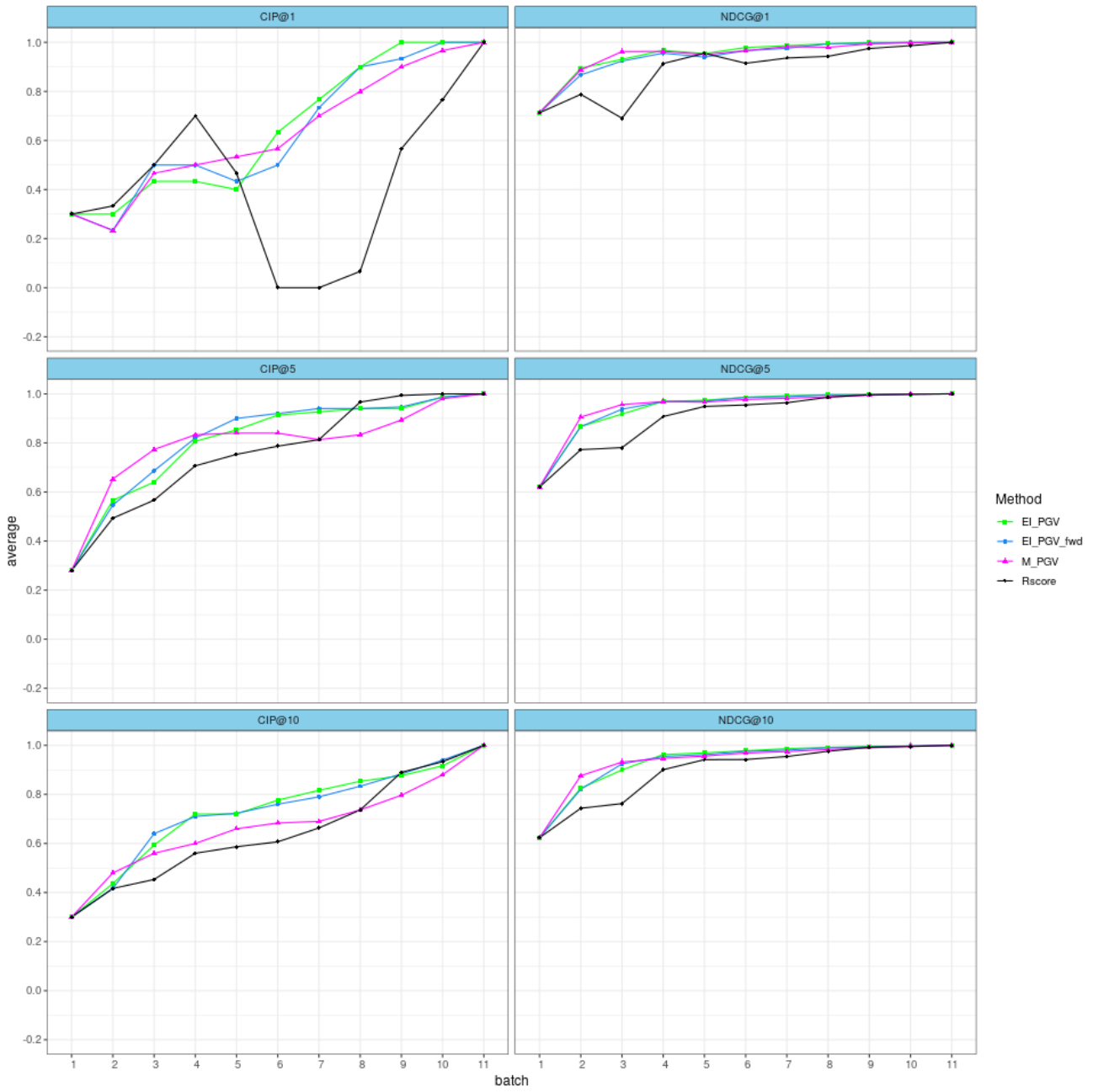
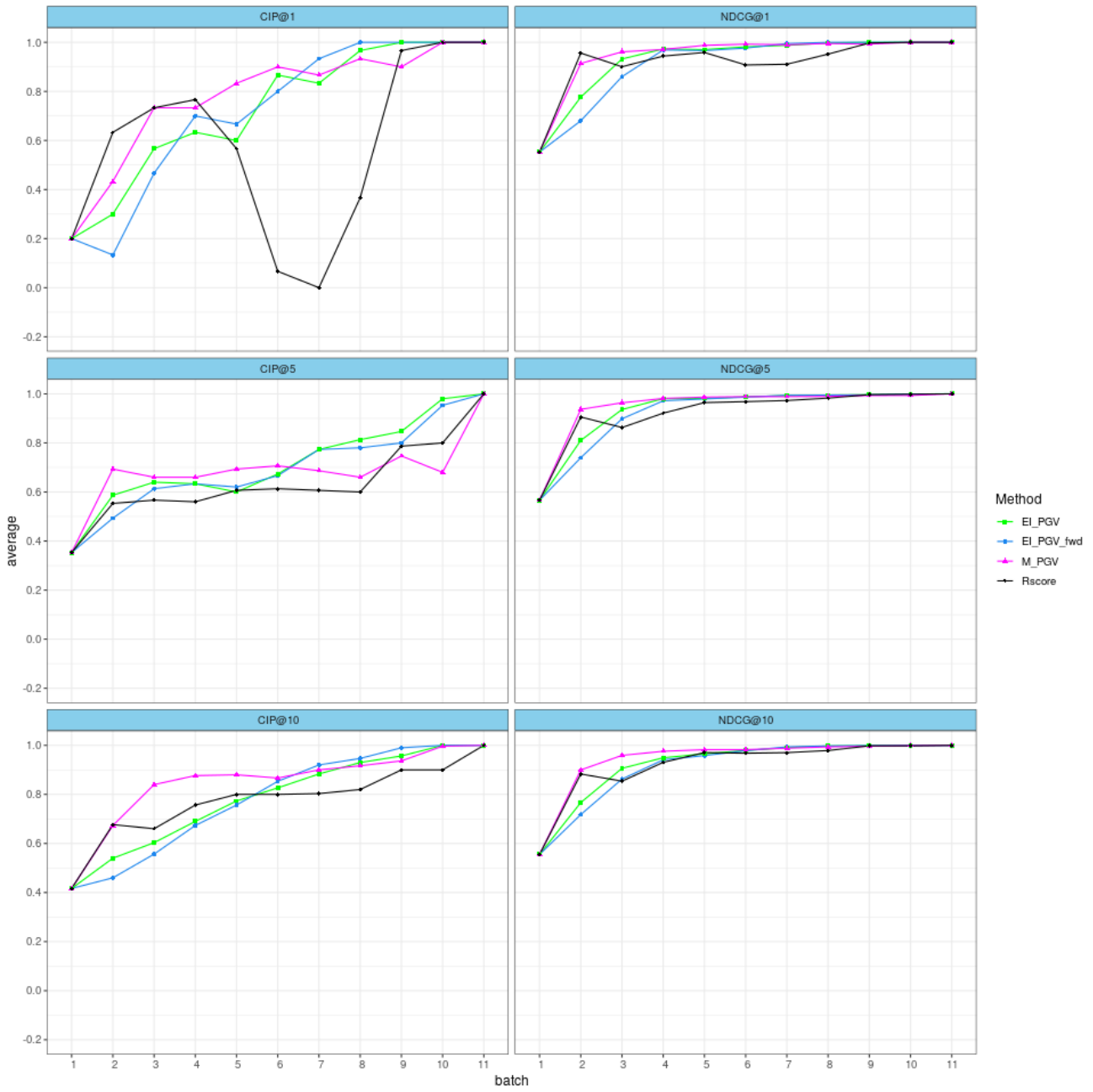


Figure 2. The assessment indices for case (ii). For the tropical rice breeding lines dataset with the batch size is set to 10, the average of correct identification proportion (CIP) and normalized discounted cumulative gain of top k individuals (NDCG@ k), which were calculated at each batch over the 30 repetitions.

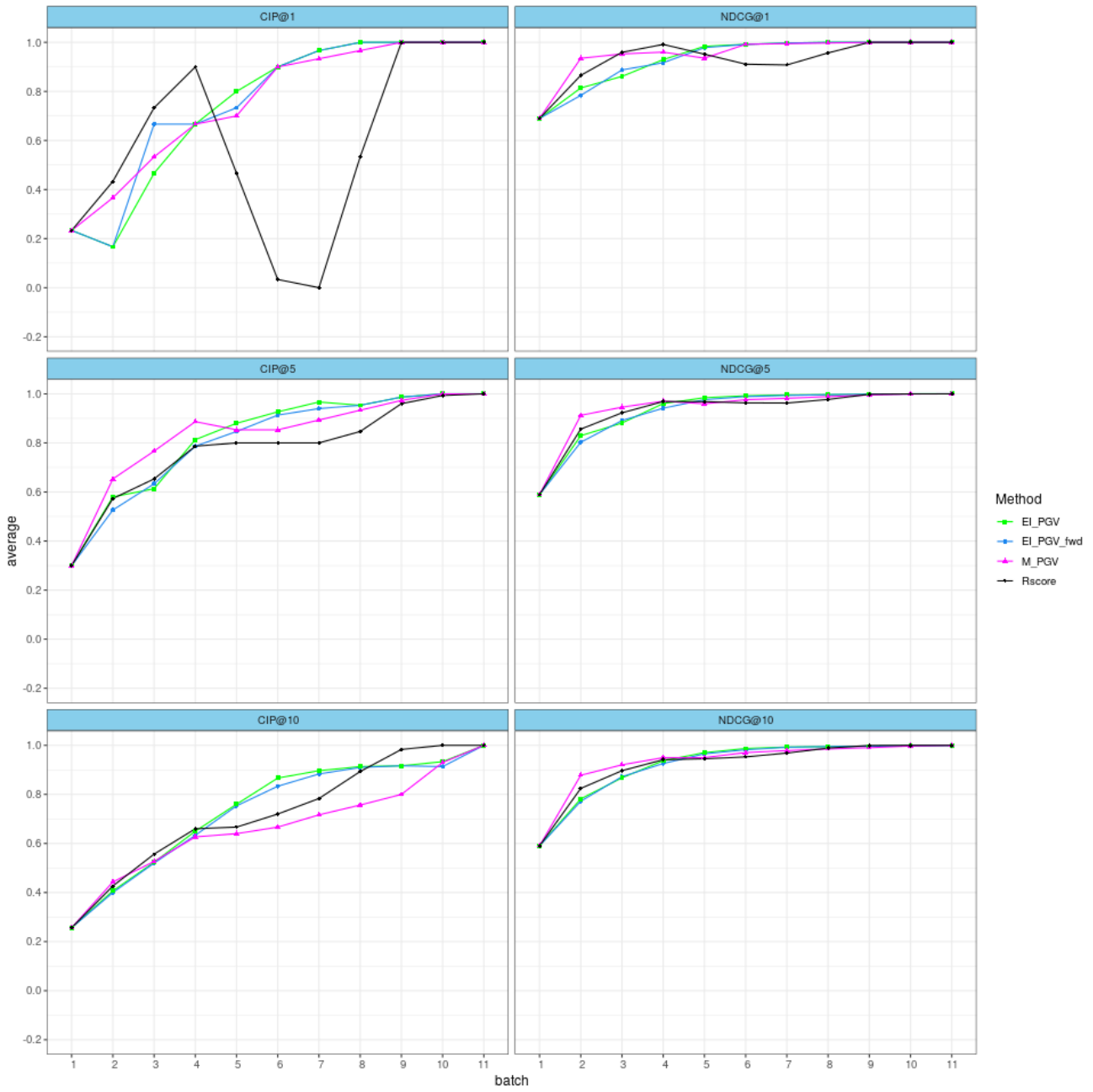
(a) Scenario 1



(b) Scenario 2



(c) Scenario 3



(d) Scenario 4

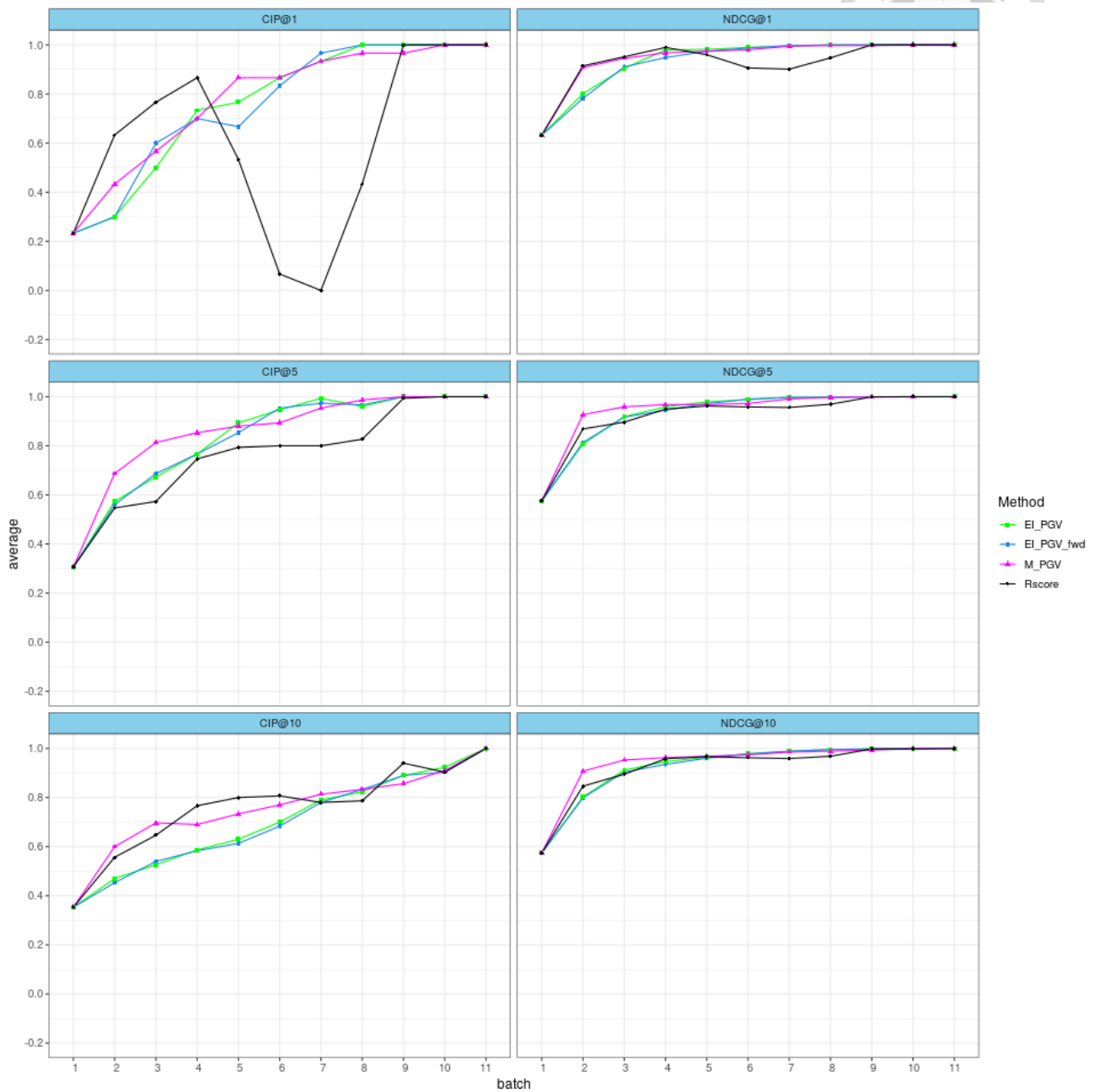
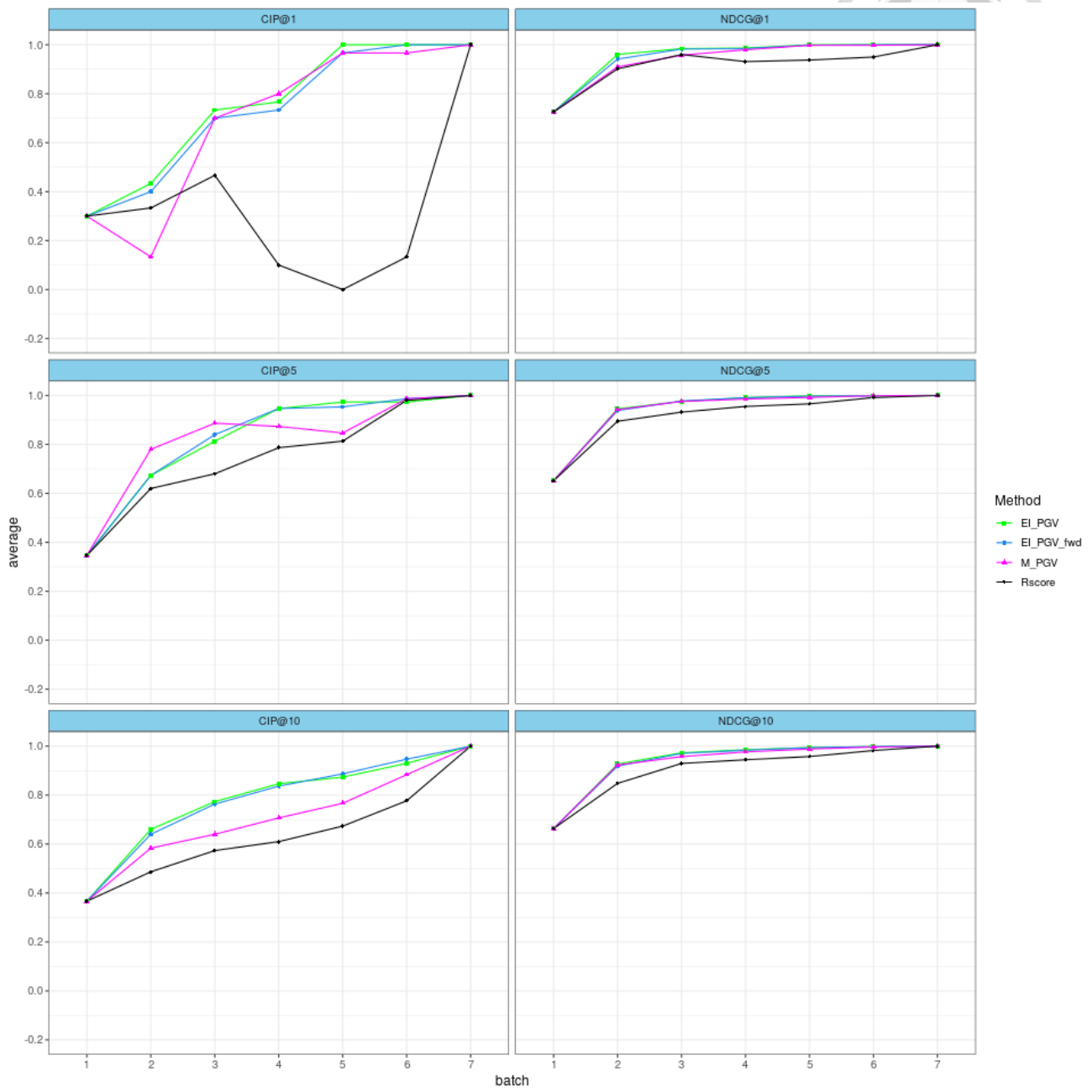
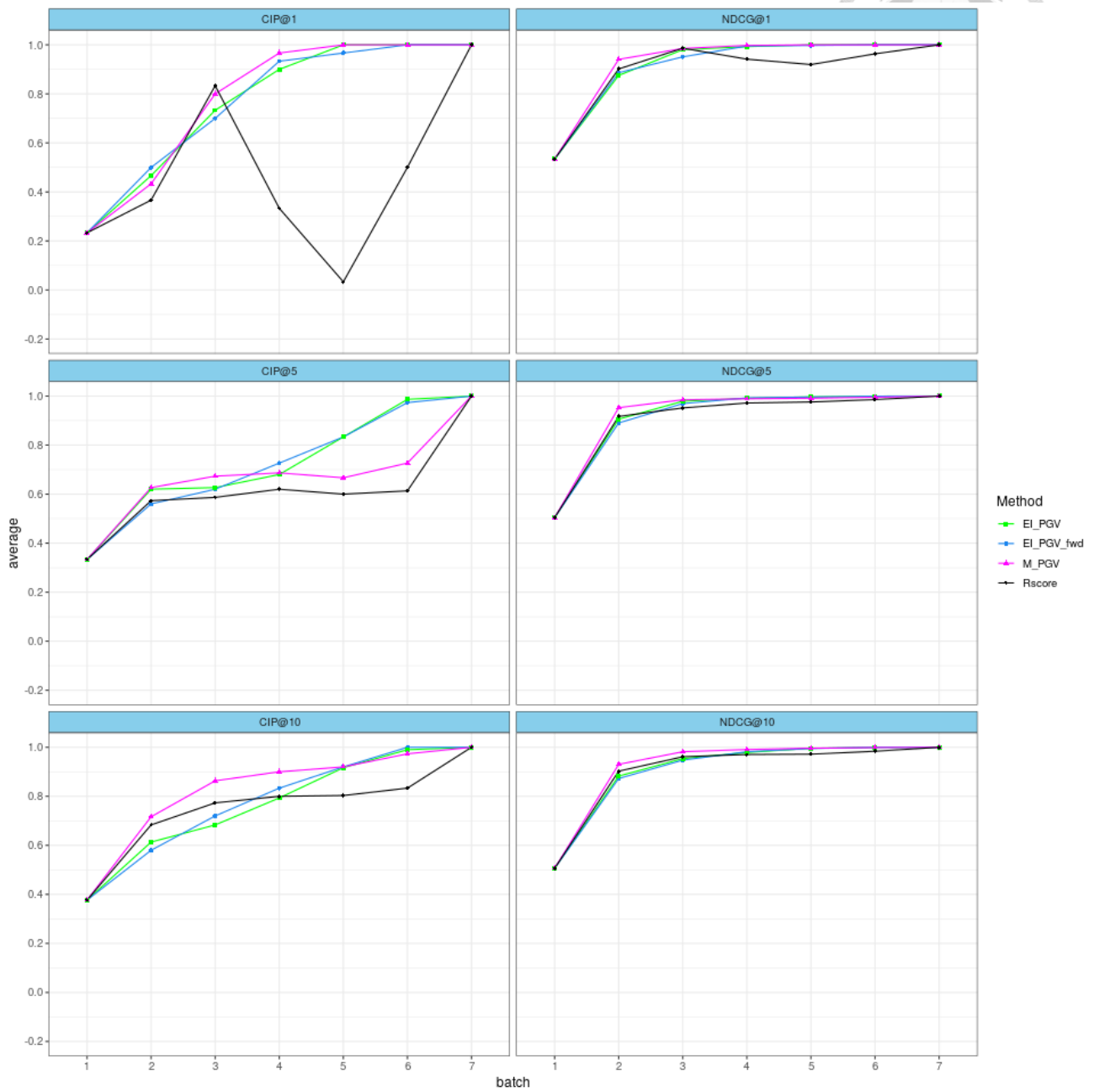


Figure 3. The assessment indices for case (iii). For the tropical rice breeding lines dataset with the batch size is set to 30, the average of correct identification proportion (CIP) and normalized discounted cumulative gain of top k individuals (NDCG@ k), which were calculated at each batch over the 30 repetitions.

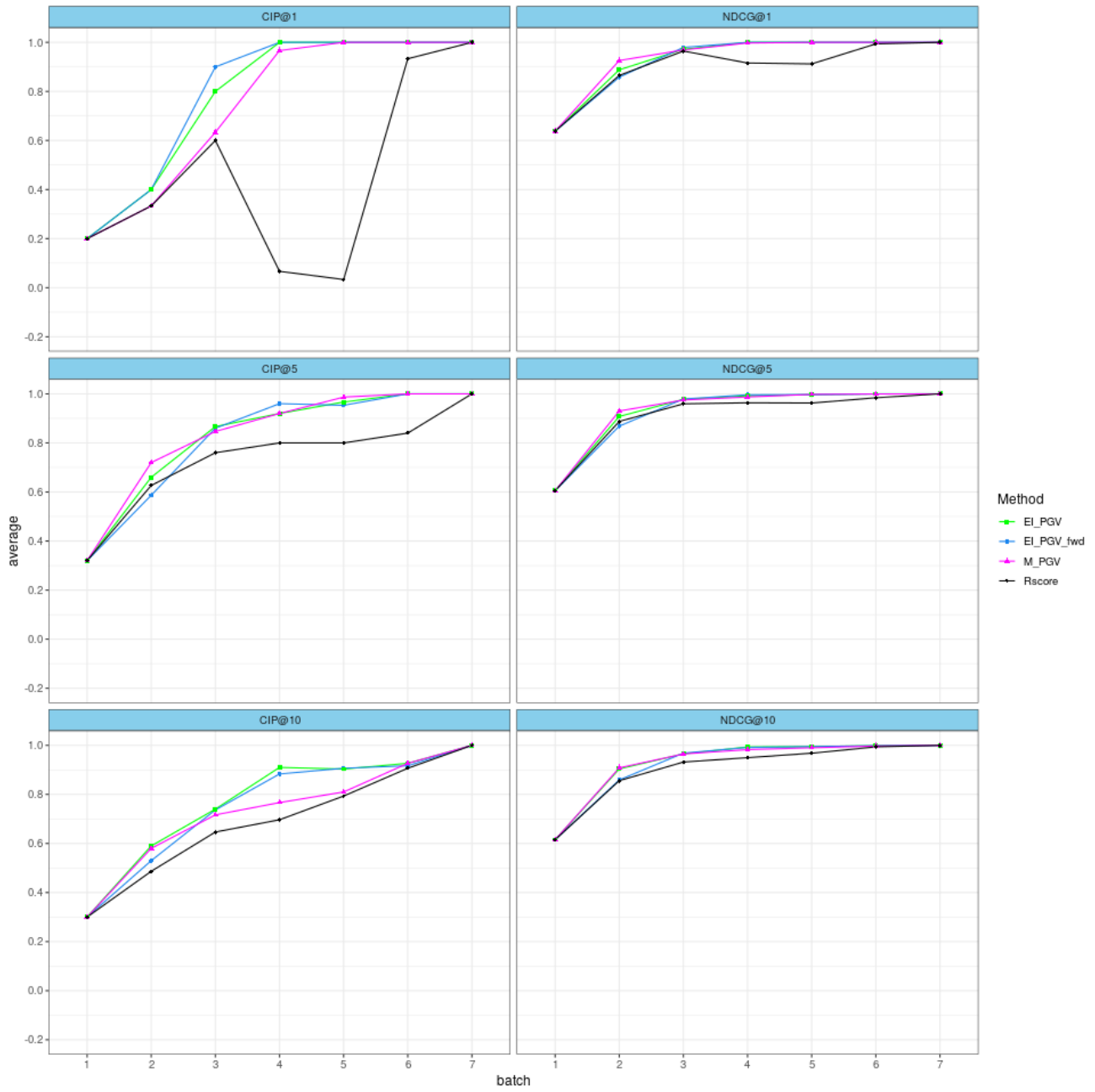
(a) Scenario 1



(b) Scenario 2



(c) Scenario 3



(d) Scenario 4

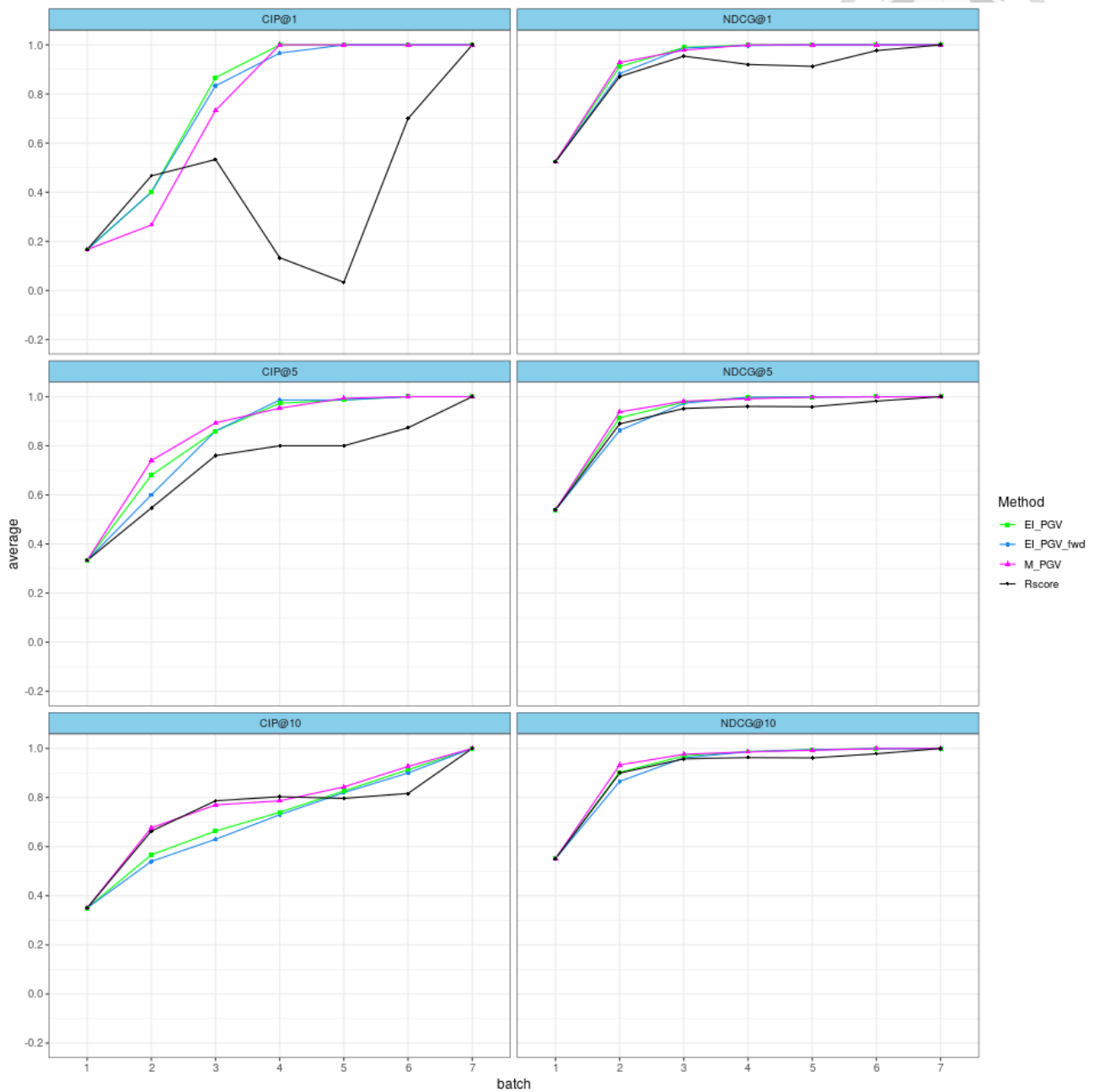


Figure 4. The assessment indices for case (iv). For the tropical rice breeding lines dataset with the batch size is set to 50, the average of correct identification proportion (CIP) and normalized discounted cumulative gain of top k individuals (NDCG@ k), which were calculated at each batch over the 30 repetitions.

(a) Scenario 1

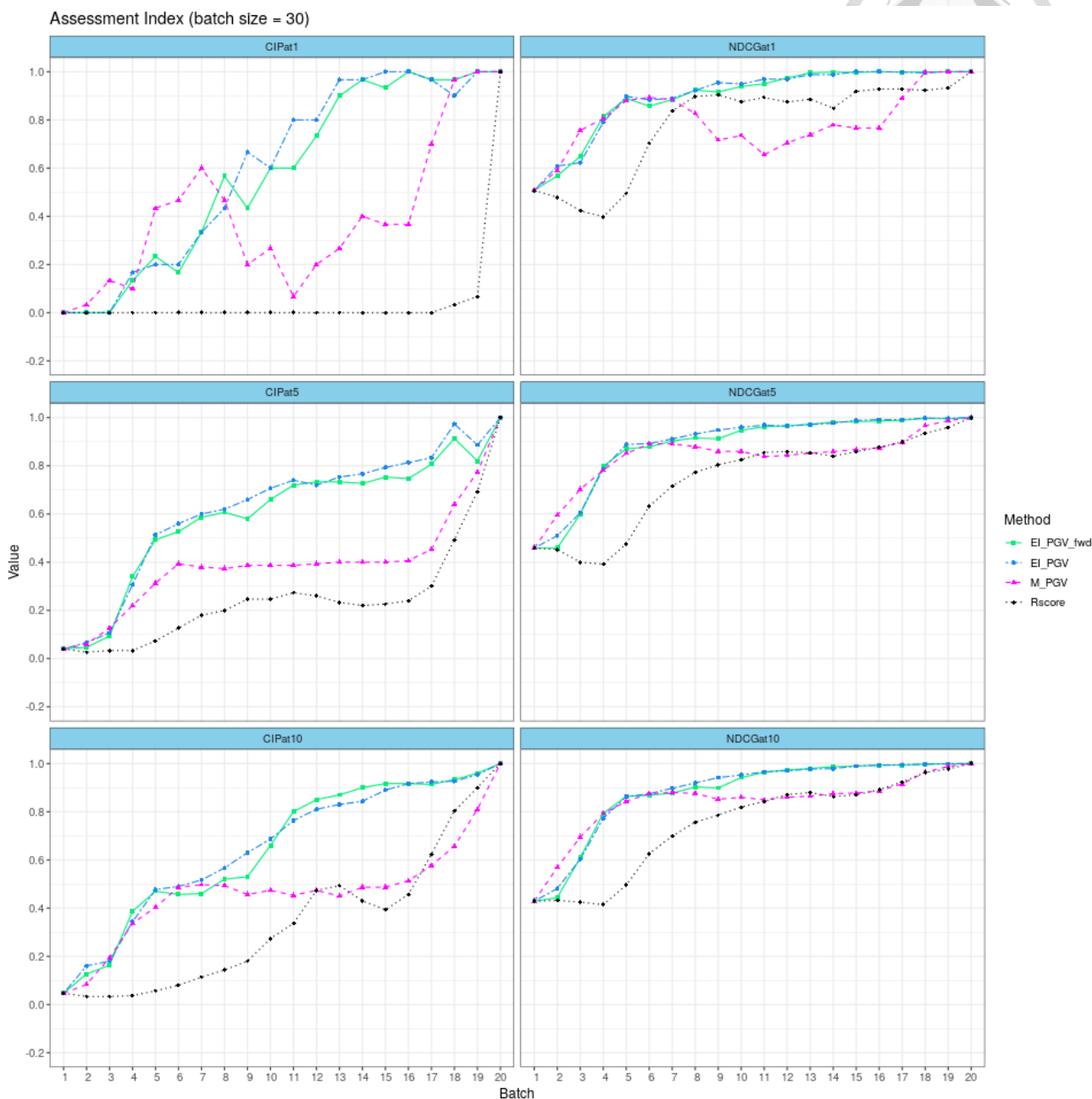
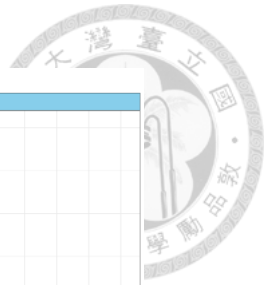
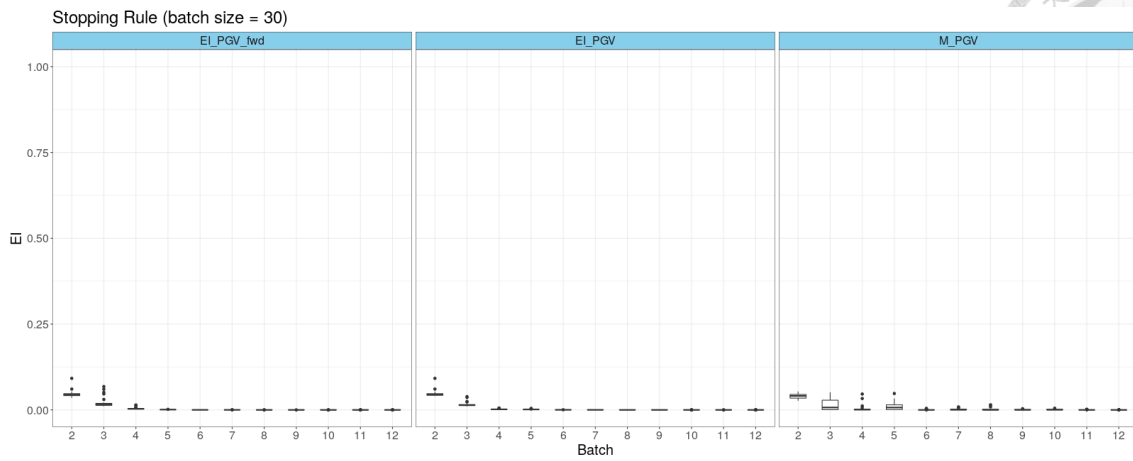


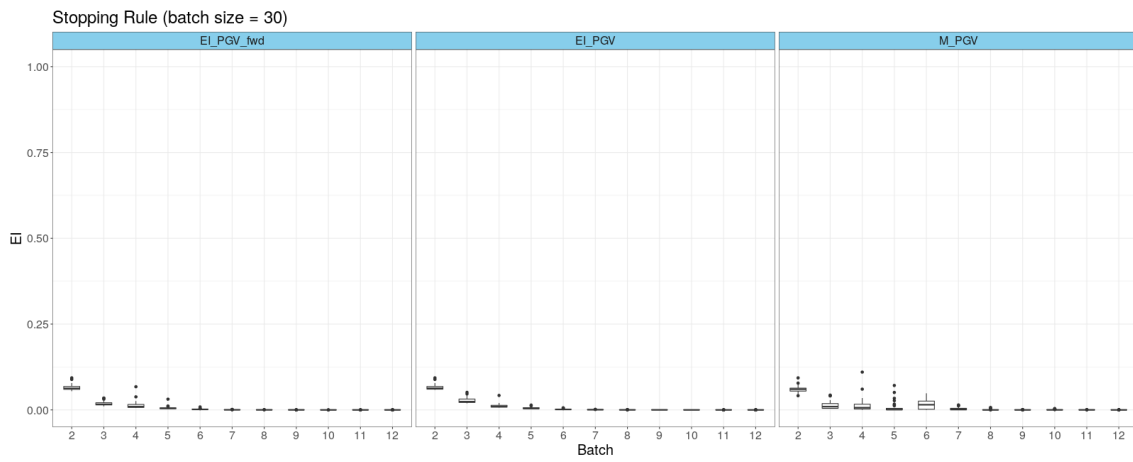
Figure 5. The assessment indices for case (v). For the wheat dataset with the batch size is set to 30, the average of correct identification proportion (CIP) and normalized discounted cumulative gain of top k individuals (NDCG@ k), which were calculated at each batch over the 30 repetitions.



(a) Scenario 1



(b) Scenario 2



(c) Scenario 3

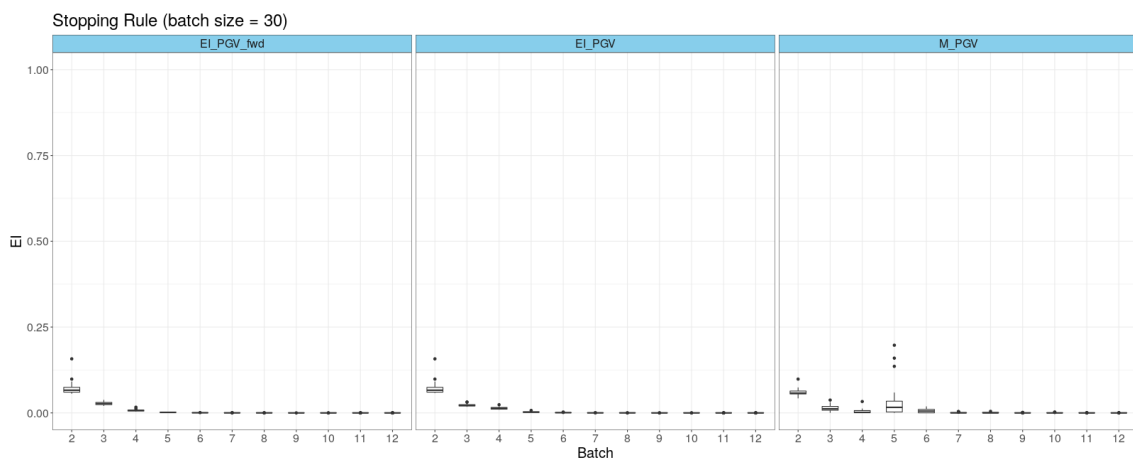
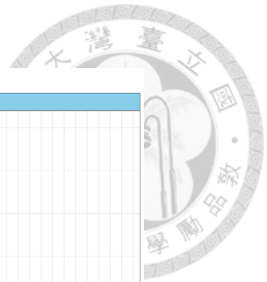
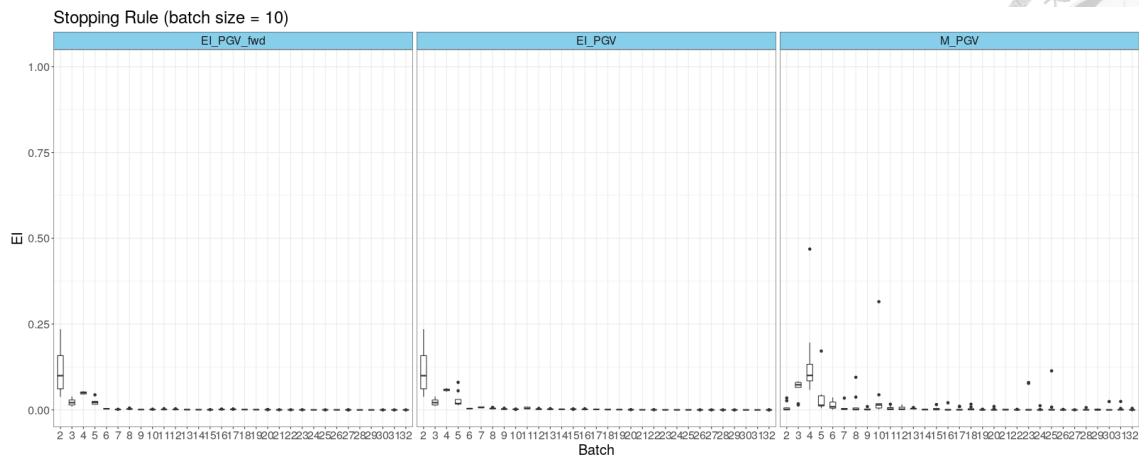


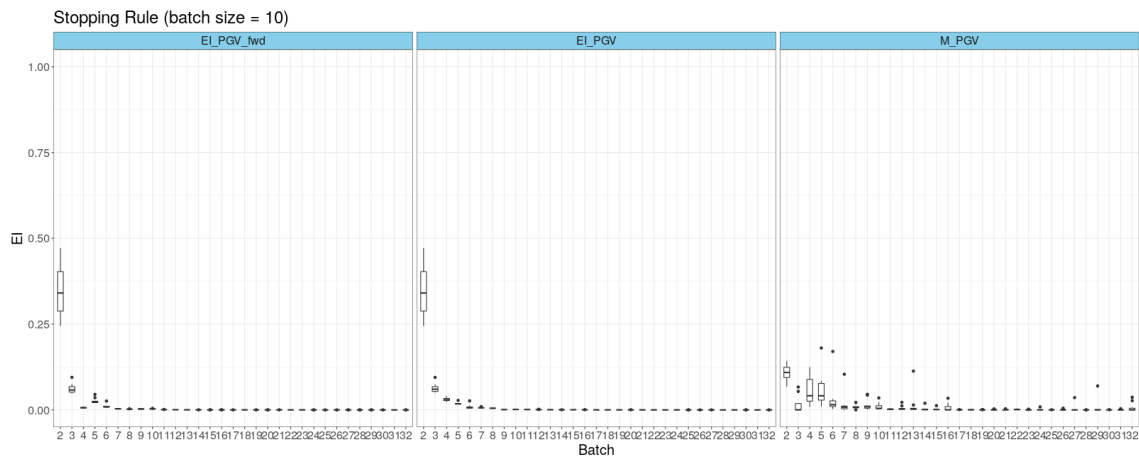
Figure 6. The stopping rule for case (i). For the 44k rice dataset with the batch size is set to 30, the expected values of training set which is determined by EI-PGV-fwd, EI-PGV and M-PGV at each batch.



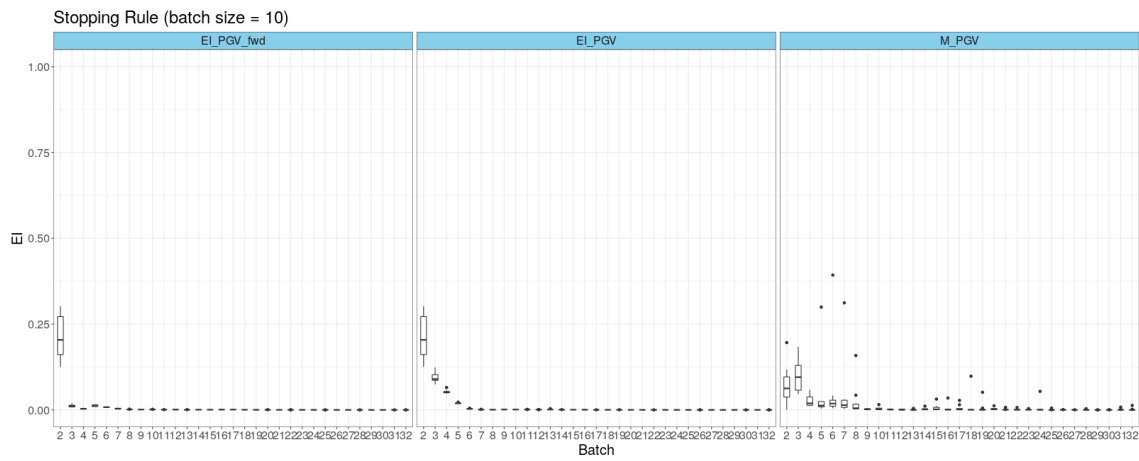
(a) Scenario 1



(b) Scenario 2



(c) Scenario 3



(d) Scenario 4

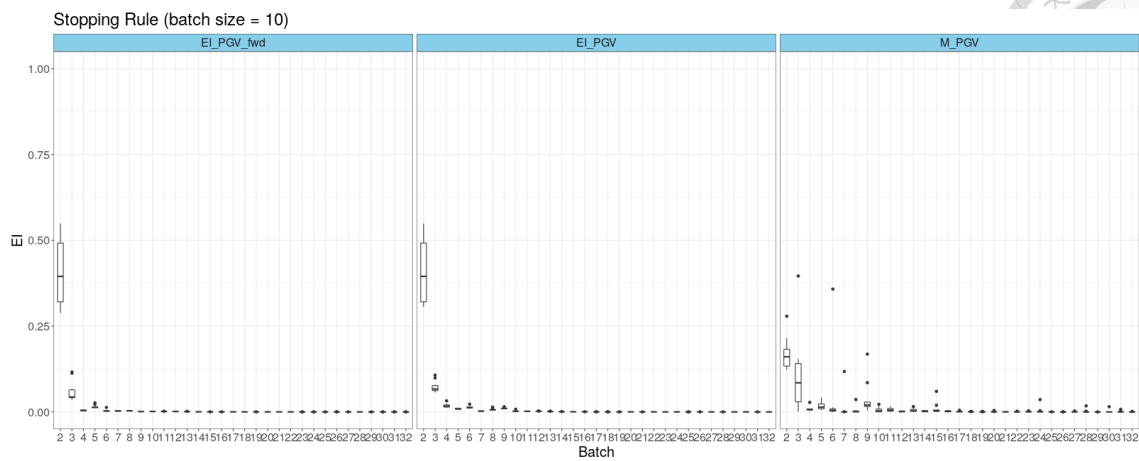


Figure 7. The stopping rule for case (ii). For the tropical rice breeding lines dataset with the batch size is set to 10, the expected values of training set which is determined by EI-PGV-fwd, EI-PGV and M-PGV at each batch.

(d) Scenario 4

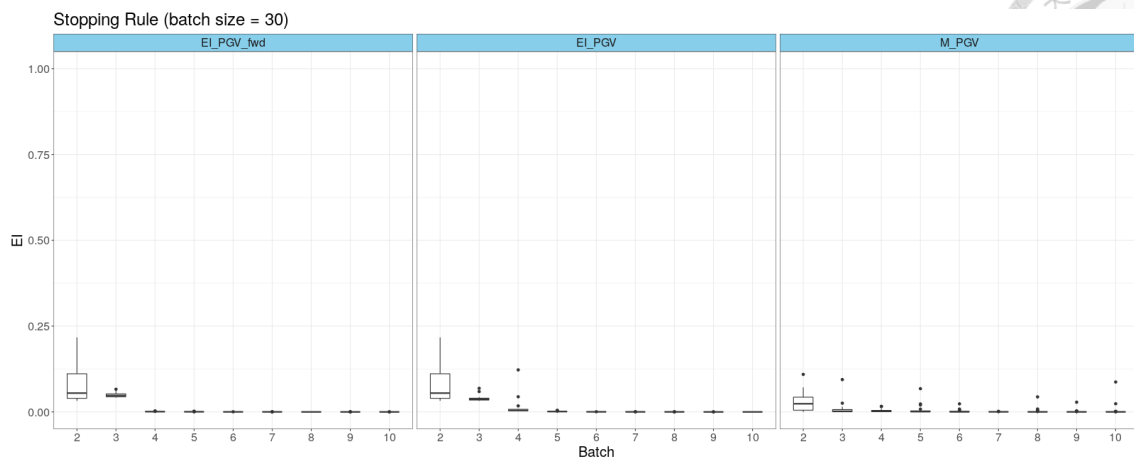
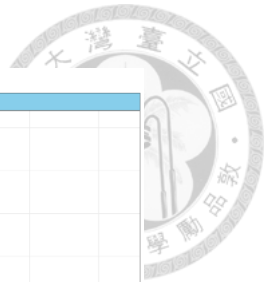
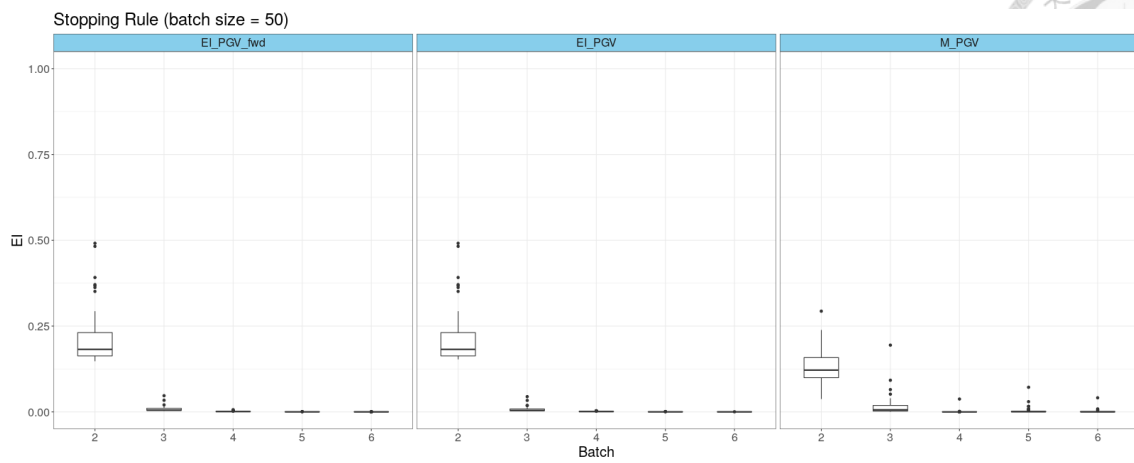


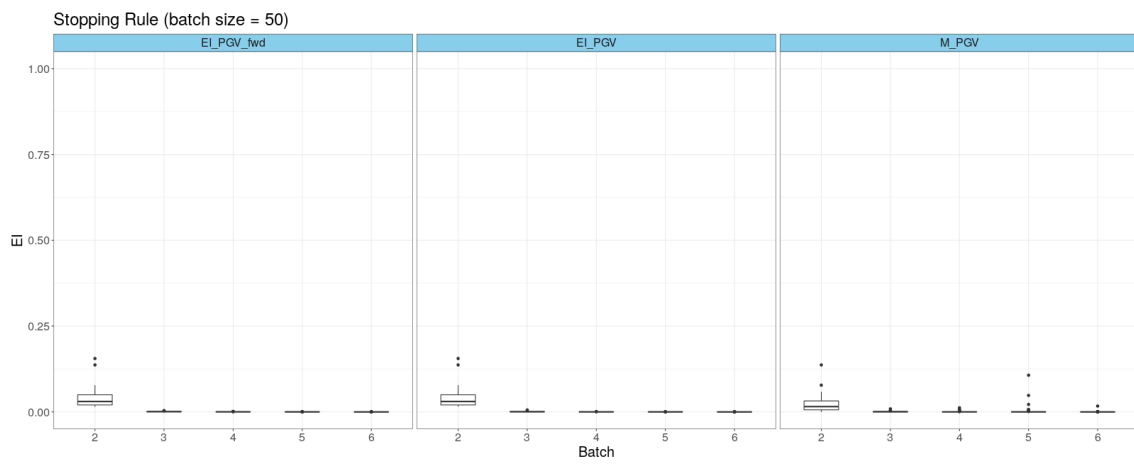
Figure 8. The stopping rule for case (iii). For the tropical rice breeding lines dataset with the batch size is set to 30, the expected values of training set which is determined by EI-PGV-fwd, EI-PGV and M-PGV at each batch.



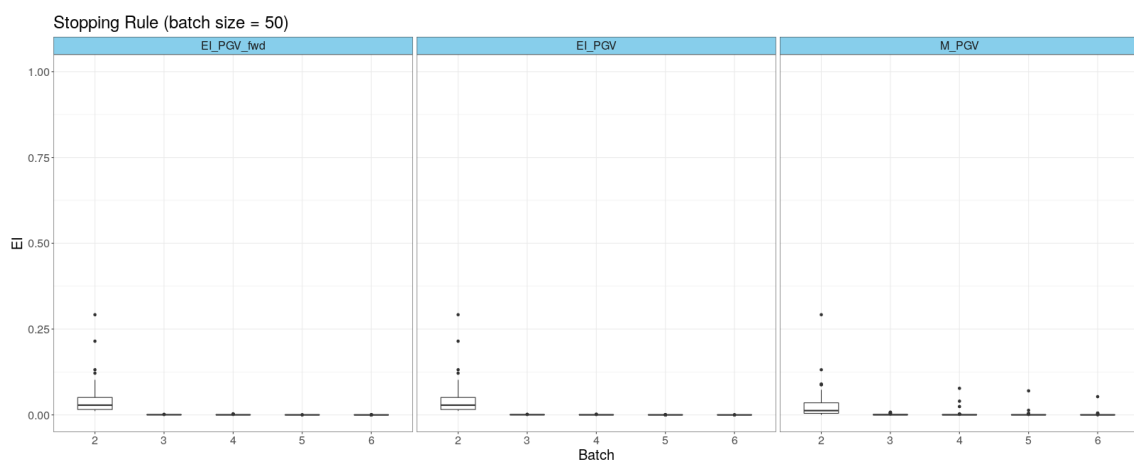
(a) Scenario 1



(b) Scenario 2



(c) Scenario 3



(d) Scenario 4

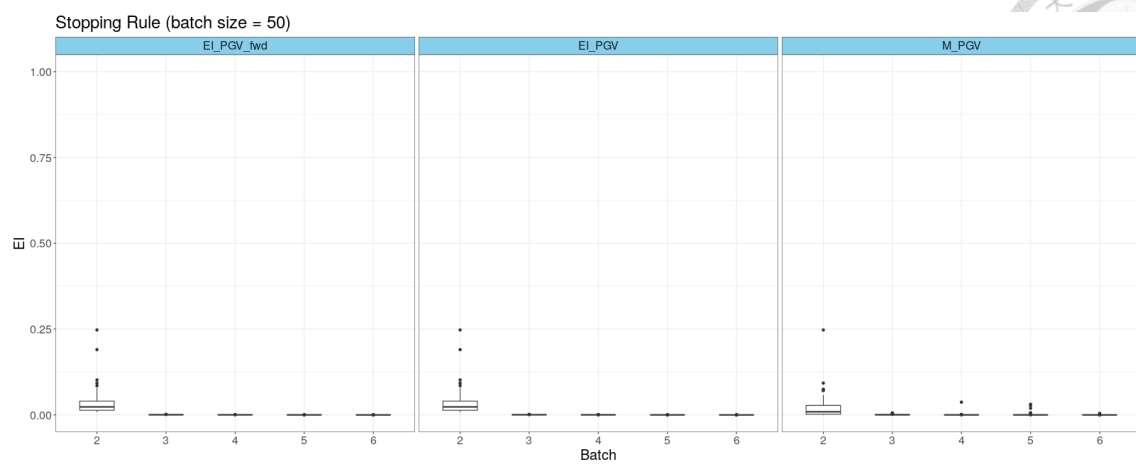


Figure 9. The stopping rule for case (iv). For the tropical rice breeding lines dataset with the batch size is set to 50, the expected values of training set which is determined by EI-PGV-fwd, EI-PGV and M-PGV at each batch.

(a) Scenario 1

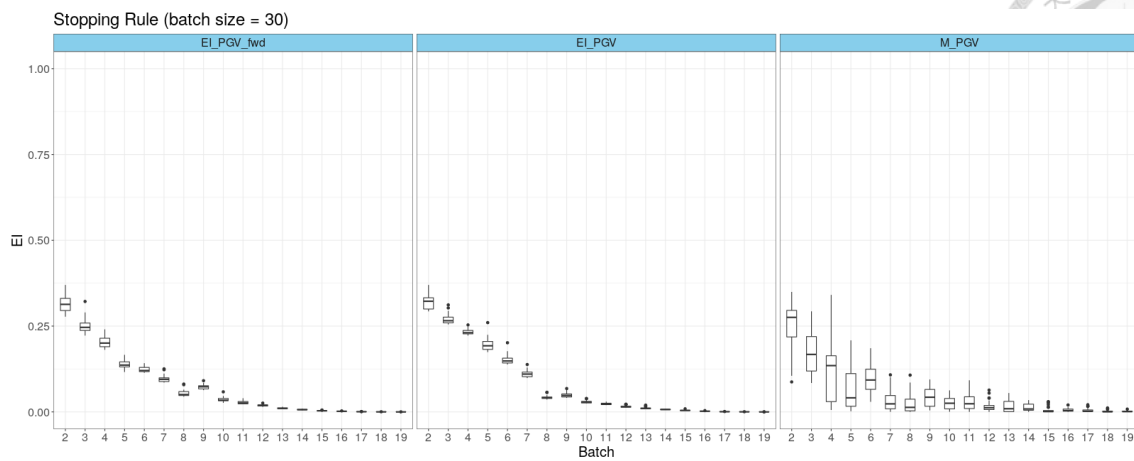


Figure 10. The stopping rule for case (v). For the tropical rice breeding lines dataset with the batch size is set to 30, the expected values of training set which is determined by EI-PGV-fwd, EI-PGV and M-PGV at each batch.