Abstract

In this Master Thesis, we have analytically derived and numerically implemented three estimators of the Prediction Divergence Criterion (Avella-Medina et al., working paper) for Model Selection within the logistic regression framework. After the validation of these estimators by means of simulations, we have performed Model Selection both when the order of the variables was known in advance and when the order was correct but decided by an already existing algorithm, namely the binary lasso (Friedman et al., 2010). Finally we have produced evidences of the good performance of two of these estimators, one derived from the L2 norm error measure and the other from the binomial deviance, respectively in highly and moderately correlated settings. They have been proven better, to the extension of the simulation study, than the defaults methods, based on 10-fold Cross Validation, currently available in the glmnet(Friedman et al., 2017) R package.

Reference


Available at:
http://archive-ouverte.unige.ch/unige:102928

Disclaimer: layout of this document may differ from the published version.
Prediction Divergence Criterion for Model Selection in the Logistic Regression

Master Thesis
University of Geneva

Cesare Miglioli

Under the Supervision of
Prof. Maria-Pia Victoria-Feser (University of Geneva)

February 1, 2018
Acknowledgments

I would like to thank first of all Prof. Maria-Pia Victoria-Feser (University of Geneva) for her incredible help and commitment in the supervision of this project. Her insightful comments and suggestions greatly improved the quality of this work. I would also like to thank Prof. Davide La Vecchia (University of Geneva) for his willingness to participate in the Jury of the Master Thesis Defense.

In the end I am happy to mention other amazing people for the fruitful discussions we had on the topic of the thesis: Guillaume Blanc, prof. Stephane Guerrier, Samuel Orso and Haotian Xu.

Of course I feel the moral need to thank my parents Euriola and Flavio *causa prima* of the birth of the Author, my brother Edoardo and all my friends and relatives.

All remaining errors, typos or inconsistencies of this work are of my own.
## Contents

1 **Introduction** .................................................. 7

2 **Analytical Derivation and Numerical Validation** .................. 11
   2.1 Analytical Formulae of the PDC in the Logistic Regression .......... 11
     2.1.1 PDC L2 Norm ............................................. 11
     2.1.2 PDC Binomial Deviance .................................. 12
     2.1.3 PDC Hamming Distance .................................. 13
   2.2 Validation of the Covariance Estimators .......................... 14
     2.2.1 L2 Norm ............................................... 15
     2.2.2 Binomial Deviance ..................................... 16
     2.2.3 Hamming Distance ..................................... 17

3 **Model Selection via the PDC with Oracle Ordering** .............. 19
   3.1 Simulation Setting ........................................... 19
   3.2 PDC Stopping Rule ......................................... 20
   3.3 Simulation Results ........................................ 22
     3.3.1 When n > p ........................................... 23
     3.3.2 When n < p ........................................... 25

4 **PDC and CV Comparison under Lasso Ordering** .................... 27
   4.1 Simulation Setting ........................................... 27
   4.2 Simulation Results ........................................ 28

5 **Conclusion** .................................................. 31

6 **References** .................................................. 32

A **Analytical Derivations** ...................................... 34
   A.1 L2 Norm .................................................. 34
   A.2 Binomial Deviance ........................................ 36
   A.3 Hamming Distance ........................................ 38
Statistical setup

In this thesis, we consider a random variable $Y$ distributed according to a parametric model $F_\beta$ conditionally on a set of fixed covariates $x = [x_1 \ldots x_p]$. We indicate a random sample (of size $n$) generated from $F_\beta$ as $y = (Y_i)_{i=1,\ldots,n}$ together with a fixed $n \times p$ full rank matrix of predictors $X$. A prediction function based on the chosen model $F_\beta$, will be noted as $\hat{y}$. We also let $y^0 = (Y_i^0)_{i=1,\ldots,n}$ indicate a random vector distributed as $y$ which can be interpreted as an out-of-sample version of $y$. Throughout this work, $E[.]$ and $E_0[.]$ denote expectations under the distribution of $Y_i|X$, respectively $Y_i^0|X$.

We consider in particular for $F_\beta$ the logistic regression model. This implies that $y = (Y_i)_{i=1,\ldots,n}$ becomes a vector of binary data where the $(Y_i)_{i=1,\ldots,n}$ are independent Bernoulli random variables:

$$Y_i = \begin{cases} 1 & \text{with probability } \pi_i \\ 0 & \text{with probability } 1 - \pi_i \end{cases}$$

with

$$\pi_i = \frac{\exp(x_i^T \beta)}{1 + \exp(x_i^T \beta)}, \quad i = 1, \ldots, n, \quad (1)$$

and $x_i^T$ is the $i$-th line of matrix of predictors $X$. When we substitute in (1) an estimator of $\beta$ (e.g. MLE), denoted by $\hat{\beta}$, we obtain the fitted probabilities $\hat{\pi}_i$.

When dealing with model selection in the logistic regression, we consider a model $M_j$ nested in a model $M_k$ where $j$ and $k$ represent the number of covariates of each model ($j < k$ by definition). Let, for $l = j, k$, the matrix $X_l$ be the $n \times l$ matrix formed by $l$ columns of matrix $X$, with line $x_i^{(l)}$ written as a $l \times 1$ vector, $\beta_l (\hat{\beta}_l)$ the corresponding (estimated) slopes and $\hat{y}_i^{(l)}$ the corresponding vector of predicted responses $\hat{y}_{i,l}$ based on the fitted probabilities $\hat{\pi}_i^{(l)}$. 
1 Introduction

Model Selection is an active research area in statistics, with important challenges and open questions. Model selection methods are demanded in many fields of application, like in Biology, Medicine, Economics, especially when dealing with high dimensional data (often \( p >> n \)).

In these settings, an exhaustive search (i.e. on all possible models) associated with a criterion for each model (to be optimized) is just unfeasible especially outside the linear model; for example with \( p = 100 \), there are \( (2^{100} - 1) \approx 1.267 \cdot 10^{30} \) possible models. For this reason, variable ordering (according to a given criterion) and stopping the search when a (possibly other) criterion is optimized, is currently one of the most relevant strategies to address the problem of dimensionality. However, it is worth citing another different but important approach to the problem, coming from the bayesian domain: Stochastic Variable Selection (George and McCulloch, 1993). There is a large body of literature on this topic (e.g. George and McCulloch, 1997; George, 2000; George and Wang, 2007) and in the recent work of George and Rocková (2014) there is an extension of these methods for the high dimensional setting.

Among the first class of methods explained above (i.e. variable ordering & criterion based stopping search), the lasso (Tibshirani, 1996) is the most popular. In the binary case (Friedman, Hastie and Tibshirani, 2010), the ordering is performed by varying the penalty criterion and the optimal model is found by means of k-fold (with \( k = 10 \) as a default value in the \texttt{glmnet} R package, 2017) cross-validation (e.g. Stone, 1974).

Hence, model selection in high dimensions for the first class of methods, combines on one hand ordering algorithms and on the other hand optimization criteria (for stopping the search). In the latter category, given an ordering rule, the most famous is k-fold cross-validation (e.g. Stone, 1974) and in the linear case Mallow’s \( C_p \) (Mallow, 1973), AIC (Akaike, 1973), BIC (Schwarz, 1978) and HQ (Hannan and Quinn, 1979). Efron (2004) proposes a covariance penalty framework that encompasses all these optimization criteria. We start from the \textit{q-class} of error measures (Efron, 1986) that allows to represent the assessed error \( Q(Y_i, \hat{Y}_i) \) for outcome \( Y_i \) given the prediction \( \hat{Y}_i \) as:

\[
Q(Y_i, \hat{Y}_i) = q(\hat{Y}_i) + \hat{q}(\hat{Y}_i)(Y_i - \hat{Y}_i) - q(Y_i)
\]

where \( q(\cdot) \) is any real-valued concave function which depends on the choice of the error measure and \( \hat{q}(\cdot) \) its first derivative. Avella-Medina, Guerrier and Victoria-Feser (working paper) generalize the \textit{q-class} making use of Bregman divergences (Bregman, 1967). We can think of Bregman divergences as the
multivariate equivalents of the members of the *q-class* of error measures proposed by Efron (1986). The squared error loss, the relative entropy, the logistic loss, the Mahalanobis distance and other well known error measures belong to this class whose members have often been used as loss functions in the context of Model Selection (e.g. Zhang, 2008; Zhang et al., 2009, 2010a). A Bregman divergence between two vectors \( \mathbf{u} \) and \( \mathbf{v} \) (of the same dimension) is defined by:

\[
D(\mathbf{u}, \mathbf{v}) = \psi(\mathbf{u}) - \psi(\mathbf{v}) - (\mathbf{u} - \mathbf{v})^T \nabla \psi(\mathbf{v})
\]

where \( \psi(\cdot) \) is the multivariate version of \( q(\cdot) \) and depends on the choice of the error measure. By analogy \( \nabla \psi(\mathbf{v}) \) is the multivariate version of \( \dot{q}(\cdot) \) and in this case represents the gradient vector of \( \psi(\cdot) \) evaluated at \( \mathbf{v} \). Within this new class, which is named *d-class* of error measures, we can define the *Prediction Divergence Criterion* (from now on PDC) which is a model selection criterion based on a divergence between the predictions of two nested models. The building block for the derivation of this criterion is Efron’s Optimism Theorem (Efron, 2004) which demonstrates that we can express the (out-of-sample) Expected Prediction Error (EPErr) for unit \( i \), for a given prediction rule \( \hat{y} \), as:

\[
EPErr_i = \mathbb{E}[\mathbb{E}_0[Q(Y_{i}^0, \hat{Y}_i)]|y] = \mathbb{E}[Q(Y_{i}, \hat{Y}_i) + \Omega_i]
\]

where \( Q(Y_{i}^0, \hat{Y}_i) \) is the assessed error defined previously for the *q-class* of error measures and \( \Omega_i = \text{cov}(\dot{q}(\hat{Y}_i), Y_i) \). Hence we can find an estimator for the EPErr to be:

\[
\hat{EPErr}_i = \frac{1}{n} \sum_{i=1}^{n} (Q(y_i, \hat{y}_i) + \hat{\text{cov}}(\dot{q}(\hat{Y}_i), Y_i))
\]

The objective of this Master Thesis is to derive and implement several PDC criteria for model selection within the logistic regression framework. Consider a model \( M_j \) nested in a model \( M_k \), define \( \hat{y}_j^0 \) as the out-of-sample prediction of the smaller model and \( \hat{y}_k \) as the in-sample prediction of the larger model, then applying *Theorem 1* of Avella-Medina et al. (working paper), which is the multivariate analog of Efron’s Optimism Theorem, we can can write that:

\[
PDC_{j,k} = \mathbb{E}[\mathbb{E}_0[D(\hat{y}_j^0, \hat{y}_k)|y]]
\]

Following the same methodology outlined above for the EPErr and making use of result (5) in Avella-Medina et al. (working paper), we obtain a natural estimator of the \( PDC_{j,k} \):

\[
\hat{PDC}_{j,k} = D(\hat{y}_j, \hat{y}_k) + tr[\hat{\text{cov}}(\dot{y}_j, \nabla \psi(\hat{y}_k))]
\]

where \( \psi(\cdot) \) and \( \nabla \psi(\hat{y}_k) \) have been already defined for the Bregman divergence.
The PDC has an intuitive interpretation: if the smaller model is not correct, the additional elements in the larger model will create differences in the predictions and therefore should be accounted for. Hence, a reasonable Model Selection procedure based on prediction divergence should first propose an ordered sequence of predictors, with the informative ones included first in the sequence, and secondly, by stopping when the PDC is minimized, ensure that non-informative predictors are not included in the selected model.
Thesis Outline

In this Master Thesis, building on the theoretical framework of the PDC explained above, three valid Bregman divergences (Bregman, 1967) will be selected in the $d$-class (i.e. multivariate extension of the $q$-class) of error measures for the specific case of the logistic regression.

In the first section, a specific analytical expression of the estimator of the PDC will be derived for the squared error loss (i.e. L2 norm), the Hamming distance (i.e. counting errors) and the binomial deviance which is twice the Kullback-Leibler divergence (Kullback-Leibler, 1951). In order to obtain the analytical formulae, three kinds of approximations for the $\hat{\text{cov}}(\hat{y}_j, \nabla \psi(\hat{y}_k))$ will be employed: a standard exponential family result (Efron 1978a, eq. 2.4), a first order Taylor expansion around the true value of the estimator and, only in the Hamming distance case, the Central Limit Theorem. These approximations will be tested in the second part of the first section by means of simulations of the covariance quantity outlined above and comparison with the analytical results.

The second section of the Thesis is devoted to the Model Selection task. A simulation study will be performed in a sparse setting with three varying parameters: the sample size $n$, the number of available explanatory variables $p$ and $\rho$ the correlation between the covariates. The simulation structure is taken from Friedman, J., Hastie,T. and Tibshirani, R. (2010) and the order of the variables is considered as known in advance. The aim of this section is to test the Model Selection properties of the three PDC estimators in a setting where the potential randomness induced by the ordering procedure is absent.

In the third section, will be presented a comparison of the Model Selection performance of the three PDC estimators with the 10-fold cross-validation (CV from now on), which is the default criterion in the R package glmnet (glmnet R package, 2017) created by Friedman, Hastie and Tibshirani. With respect to the ordering rule, the one provided by the lasso for the logistic regression (Friedman et al., 2010) of the same glmnet package is the natural choice. The simulation setting of this section will be chosen in order to achieve a correct ordering (i.e. all the active regressors are placed before the inactive ones) from the binary lasso algorithm and thus guaranteeing a meaningful comparison.

In the Conclusion section, the limitations and the future research directions of this work will be discussed.
Analytical Derivation and Numerical Validation

This section is divided in two parts: in the first one we present the analytical derivations of the estimator of the PDC for the squared error loss (i.e. L2 norm), the binomial deviance and the Hamming distance (i.e. counting errors) in the logistic regression case. In the second part the theoretical quantities found in the previous part are verified by means of simulation techniques.

2.1 Analytical Formulae of the PDC in the Logistic Regression

As already mentioned in the previous section, if we consider a model $M_j$ nested in a model $M_k$, and consider $\hat{y}_l$, $l = j, k$ the predictions computed at respectively the smaller and larger models, we have:

$$PDC_{j,k} = D(\hat{y}_j, \hat{y}_k) + tr[\hat{cov}(\hat{y}_j, \nabla \psi(\hat{y}_k))]$$

(2)

where $\psi(\cdot)$ is a scalar function that is related to the choice of the error measure and $\nabla \psi(\hat{y}_k)$ represents the gradient vector of $\psi(\cdot)$ evaluated at $\hat{y}_k$. We shall refer to the first and second terms of (2) as the apparent divergence and the divergence optimism, respectively.

For the logistic regression framework, we have to adapt the results presented in Table 1 at page 462 of Efron’s 1986 paper to the multivariate case scenario outlined in (2). We first define $\psi(\cdot)$ for the L2 norm, the binomial deviance and the Hamming distance, and then derive a (closed form) expression for $\hat{cov}(\hat{y}_j, \nabla \psi(\hat{y}_k))$ for every error measure. In order to preserve the simplicity of the exposition, the derivation of the latter is provided in the Appendix A at the end of this document.

2.1.1 PDC L2 Norm

In the specific case of the L2 error measure, the apparent divergence $D(\hat{y}_j, \hat{y}_k)$ becomes $\|\hat{y}_j - \hat{y}_k\|_2^2$. This implies that $\psi(\hat{y}_k) = \hat{y}_k^T \hat{y}_k$ and $\nabla \psi(\hat{y}_k) = 2 \hat{y}_k$, so that $\hat{cov}(\hat{y}_j, \nabla \psi(\hat{y}_k)) = 2 \hat{cov}(\hat{y}_j, \hat{y}_k)$.

To derive $\hat{cov}(\hat{y}_j, \hat{y}_k)$, we use an approximation proposed by Efron (Efron 1978a, eq. 2.4) for the exponential family and a first order Taylor expansion. The derivations are provided in Section A.1. The resulting estimator is then:

$$PDC_{j,k} = \sum_{i=1}^{n} (\hat{y}_{i,j} - \hat{y}_{i,k})^2 + 2 \left[ \sum_{i=1}^{n} \hat{g}(\xi_i^{(j)}T \hat{\Sigma}_{j}^{-1} \hat{\theta}_j) \hat{g}(\xi_i^{(k)}T \hat{\Sigma}_{k}^{-1} \hat{\theta}_k) \xi_i^{(j)}T \hat{\Sigma}_{j}^{-1} \hat{\Sigma}_{j,k} \hat{\Sigma}_{k}^{-1} \xi_i^{(k)} \right]$$
where, for \( l = j, k \), \( \tilde{\theta}_l = \frac{\sum_{i=1}^n x_i(l) \hat{\pi}_i(l)}{\sum_{i=1}^n \hat{\pi}_i(l) (1 - \hat{\pi}_i(l)) x_i(l) x_i(l)^T} \),
\( \hat{\Sigma}_{jk} = \frac{\sum_{i=1}^n \hat{\pi}_i(1 - \hat{\pi}_i) x_i(j) x_i(k)^T}{\sum_{i=1}^n \hat{\pi}_i(1 - \hat{\pi}_i) x_i(j) x_i(k)^T} \) and \( \hat{\Sigma}_{jk} = \frac{\exp(\lambda)}{1+\exp(\lambda)} \left( 1 - \frac{\exp(\lambda)}{1+\exp(\lambda)} \right) \).

We remember also that, as stated in Section A.1 and along the lines of the derivations presented in Avella-Medina et al. (working paper), we consistently use \( \hat{\pi}_i(j) \) whenever \( \hat{\pi}_i \) appears in a given formula.

### 2.1.2 PDC Binomial Deviance

For the binomial deviance error measure, the apparent divergence becomes:

\[
D(\hat{y}_j, \hat{y}_k) = -\frac{2}{n} \sum_{i=1}^n \left\{ \hat{y}_{i,j} \left[ \log \left( \frac{\hat{y}_{i,k}(1 - \hat{y}_{i,j})}{(1 - \hat{y}_{i,k})\hat{y}_{i,j}} \right) \right] + \log \left( \frac{1 - \hat{y}_{i,k}}{1 - \hat{y}_{i,j}} \right) \right\}
\]

(3) is obtained by setting \( \psi(\hat{y}_k) = 2 \left[ \hat{y}_k^T \log(\hat{y}_k) + (1 - \hat{y}_k)^T \log(1 - \hat{y}_k) \right] \) which implies that \( \nabla \psi(\hat{y}_k) = 2 \log(\text{diag}(\hat{y}_k(1 - \hat{y}_k)^{-T})) \). Hence, the divergence optimism is

\[
\text{tr}[\text{cov}(\hat{y}_j, \nabla \psi(\hat{y}_k))] = 2 \sum_{i=1}^n \text{cov} \left( \hat{y}_{i,j}, \log \left( \frac{\hat{y}_{i,k}}{1 - \hat{y}_{i,k}} \right) \right)
\]

(4)

To produce a closed form expression for the covariance quantity of equation (4), we use the same approximations of the L2 case. The derivations are provided in Section A.2. The resulting estimator is then:

\[
PDC_{j,k} = n^{-1} \sum_{i=1}^n -2 \left\{ \hat{y}_{i,j} \left[ \log \left( \frac{\hat{y}_{i,k}(1 - \hat{y}_{i,j})}{(1 - \hat{y}_{i,k})\hat{y}_{i,j}} \right) \right] + \log \left( \frac{1 - \hat{y}_{i,k}}{1 - \hat{y}_{i,j}} \right) \right\} + 2 \left[ \sum_{i=1}^n \hat{g}(\hat{x}_i(j)^T \hat{\Sigma}_j^{-1} \hat{\theta}_j) \hat{x}_i(j)^T \hat{\Sigma}_j^{-1} \hat{\Sigma}_{j,k} \hat{\Sigma}_k^{-1} \hat{x}_i(k) \right]
\]

where all the above quantities have been already defined in Section 2.1.1.
2.1.3 PDC Hamming Distance

The Hamming distance, also known as "counting errors" in logistic regression, is calculated summing up the differences between binary (i.e. 0 and 1) strings. In this specific case, \( \hat{y}_{i,j} \) and \( \hat{y}_{i,k} \) are indicator variables whose values depend on the choice of the cutoff point \( C_0 \in (0, 1) \). This choice can be adapted to account for the specific characteristics of the problem at hand: throughout this work we have fixed \( C_0 = 0.5 \) (common value). Thus the apparent divergence for the Hamming distance becomes:

\[
D(\hat{y}_j, \hat{y}_k) = \frac{1}{n} \sum_{i=1}^{n} \{ \hat{y}_{i,j} \neq \hat{y}_{i,k} \}
\]  

(5)

(5) is obtained by setting \( \psi(\hat{y}_k) = \hat{y}_k^T \hat{y}_k - \hat{y}_k \) which implies \( \nabla \psi(\hat{y}_k) = 2\hat{y}_k - 1 \), we can find \( \text{cov}(\hat{y}_j, \nabla \psi(\hat{y}_k)) = 2 \text{cov}(\hat{y}_j, \hat{y}_k) \).

To derive \( \text{cov}(\hat{y}_j, \hat{y}_k) \), we use again Efron’s approximation (Efron 1978a, eq. 2.4) and a Normal approximation driven by Lyapunov’s CLT (e.g. Efron 1986, (4.16)). The derivations are provided in Section A.3. The resulting estimator is then:

\[
PDC_{j,k} = n^{-1} \sum_{i=1}^{n} \{ \hat{y}_{i,j} \neq \hat{y}_{i,k} \} + 2 \sum_{i=1}^{n} \{ \Phi \text{d}_i [\hat{c}_{i,j}, \hat{c}_{i,j}] - \Phi \tilde{\text{d}}_i [\hat{c}_{i,j}, \hat{c}_{i,j}] \}
\]

Where \( \hat{c}_{i,j} = \log \left( \frac{C_0}{1-C_0} \right) - x_i^{(j)} \hat{\beta}_j \) and both \( \text{d}_i \) and \( \tilde{\text{d}}_i \) are the variance-covariance matrices of the two bivariate Normal random variables whose CDFs appear in the above formula.

In particular:

\[
\text{d}_i = \begin{bmatrix} \hat{d}_{ij} & \hat{d}_{ijk} \\ \hat{d}_{ikj} & \hat{d}_{ik} \end{bmatrix} \quad \text{and} \quad \tilde{\text{d}}_i = \begin{bmatrix} \hat{d}_{ij} & 0 \\ 0 & \hat{d}_{ik} \end{bmatrix}
\]

Where \( \hat{d}_{il} = x_i^{(l)}(\hat{\Sigma}_l^{-1})x_i^{(l)}(l) \) for \( l = j, k \) and \( \hat{d}_{ijk} = \hat{d}_{ikj} = x_i^{(j)}(\hat{\Sigma}_j^{-1})\hat{\Sigma}_{j,k}\hat{\Sigma}_k^{-1}x_i^{(k)} \).
2.2 Validation of the Covariance Estimators

In this section, we will evaluate numerically the performance of the three covariance estimators derived previously. For this purpose we will focus on the divergence optimism which is $\text{tr}[\hat{\text{cov}}(\hat{y}_j, \nabla \psi(\hat{y}_k))]$.

The key driver for the performance of the estimators is the accuracy of Efron’s approximation $\hat{\beta} \approx \Sigma^{-1} \hat{\theta}$ (Efron 1978a, eq. 2.4), where $\hat{\theta} = \sum_{i=1}^{n} x_i y_i$ and $\Sigma = \sum_{i=1}^{n} \pi_i (1 - \pi_i) x_i x_i^T$. This approximation holds only locally. Our simulation exercise has led to the conclusion that the approximation is relatively accurate when the majority of the fitted probabilities (i.e. $\hat{\pi}_i$) stays in a specific range that has been numerically evaluated as $[0.2, 0.8]$. The implication in real settings is that when the majority of outcomes are either only 0 or 1 (i.e. corresponding to fitted probabilities near 0 or 1) then the approximation is not suitable, and hence probably as well the PDC estimator for model selection. But at the same time, the exercise of predicting outcomes when most of them are identical, can be considered as a situation where model selection methods are not necessarily appropriate.

In the next three Sections we present evidences of the accuracy of the divergence optimism estimator by means of simulations for the three divergences. For the simulation exercise, we have generated a $1000 \times 4$ matrix $X$ of covariates from a standard multivariate normal distribution (MVN) with covariance set to the identity matrix. We set $\beta = [-0.4 0.3 0.12 0]$ which produces fitted values in $[0.2, 0.8]$ interval. This choice of $\beta$ implies that the true model $M_j = M_3$ while the model $M_k$ is set to $M_4$. Then the matrix $X$ coincides with the matrix $X_k$ while $X_j$ is obtained omitting the last column of $X$. Thus we evaluate the divergence optimism estimator for the specific case of the $PDC_{j,k} = PDC_{3,4}$ which enables us to keep track of the precision of Efron’s approximation since $\hat{\beta}_j \approx \Sigma^{-1} \hat{\theta}_j$ is a direct estimator of $\beta_j = [-0.4 0.3 0.12]$. We found, out of 100 simulations, that on average, $\hat{\beta}_j = [-0.45 0.35 0.14]$.

In each of the following Sections, in order to assess the accuracy of the specific divergence optimism estimator, we compare it to the mean of the simulated divergence optimism (from now on s.d.o.). For each of the 100 simulations performed in the setting outlined above, the s.d.o. is obtained as the sum of 1000 simulated estimated sample covariances whose expressions are specific to each error measure. Then each divergence optimism estimator derived in section (2.1.1), (2.1.2) and (2.1.3) is compared with the mean of the 100 s.d.o. values previously found.
2.2.1 L2 Norm

In the L2 case, the divergence optimism is $2 \text{tr} [\hat{\text{cov}}(\hat{y}_j, \hat{y}_k)]$. Thus the simulated divergence optimism becomes:

$$2 \sum_{i=1}^{n} \hat{\text{cov}}(\hat{\pi}_i^{(j)}, \hat{\pi}_i^{(k)})$$  \hspace{1cm} (6)

where $\hat{\pi}_i^{(l)}$ is the vector of fitted probabilities derived from the logistic regression on model $M_l, l = j, k$, with $j = 3, k = 4$ and $n = 1000$.

In Table 1 are provided the summary statistics over the 100 simulations of the L2 norm simulated divergence optimism (see (6)) under the setting explained in Section 2.2.

<table>
<thead>
<tr>
<th>Optimism</th>
<th>Min</th>
<th>1st Quart</th>
<th>Median</th>
<th>Mean</th>
<th>3rd Quart</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simulated</td>
<td>1.289</td>
<td>1.362</td>
<td>1.383</td>
<td><strong>1.382</strong></td>
<td>1.405</td>
<td>1.466</td>
</tr>
</tbody>
</table>

Table 1: Summary statistics for the L2 norm simulated divergence optimism (see (6)) under the simulation setting explained in Section 2.2

As explained in the end of Section 2.2, we compare the mean of the simulated divergence optimism (in bold in Table 1) to the value of the L2 divergence optimism estimator derived in Section (2.1.1) as:

$$2 \sum_{i=1}^{n} \hat{g}(x_i^{(j)^T} \hat{\Sigma}_j^{-1} \hat{\theta}_j)\hat{g}(x_i^{(k)^T} \hat{\Sigma}_k^{-1} \hat{\theta}_k)x_i^{(j)^T} \hat{\Sigma}_j^{-1} \hat{\Sigma}_{j,k} \hat{\Sigma}_k^{-1} \hat{\Sigma}_{k}^{-1} x_i^{(k)}$$  \hspace{1cm} (7)

where in our simulation exercise $j = 3, k = 4, n = 1000$.

Out of 100 simulations, the average of the values of (7) found in the simulations is: **1.313 (0.0017)**. Thus in line with **1.382**, the mean of the s.d.o, and moreover we observe a good accuracy proved by the limited standard deviation.

In the end we can conclude that, in the framework outlined in Section 2.2 and to the extension of this simulation study, the analytical estimator for the L2 divergence optimism is unbiased and that the approximations employed in the derivation of the formula are working well.
2.2.2 Binomial Deviance

The divergence optimism is given in (4) to be $2 \sum_{i=1}^{n} \text{cov} \left( \hat{y}_{i,j}, \log \left( \frac{\hat{y}_{i,k}}{1-\hat{y}_{i,k}} \right) \right)$. Thus the simulated divergence optimism becomes:

$$2 \sum_{i=1}^{n} \text{cov}(\hat{\pi}_{i}^{(j)}, \hat{\pi}_{i}^{(k)})$$

where $\hat{\pi}_{i}^{(k)} = \log(\text{diag}(\hat{\pi}_{i}^{(k)}(1-\hat{\pi}_{i}^{(k)})^{T}))$ and where $\hat{\pi}_{i}^{(l)}$ is the vector of fitted probabilities derived from the logistic regression on model $M_l$ with $l = j, k$. Clearly in our simulation exercise $j = 3, k = 4$ and $n = 1000$.

In Table 2 are provided the summary statistics over the 100 simulations of the binomial deviance simulated divergence optimism (see (8)) under the setting explained in Section 2.2.

<table>
<thead>
<tr>
<th>Optimism</th>
<th>Min</th>
<th>1st Quart</th>
<th>Median</th>
<th>Mean</th>
<th>3rd Quart</th>
<th>Max</th>
</tr>
</thead>
</table>

Table 2: Summary statistics for the binomial deviance simulated divergence optimism (see (8)) under the simulation setting explained in Section 2.2

Now we compare the mean of the simulated divergence optimism (in bold in Table 2) to the value of the binomial deviance divergence optimism estimator derived in Section 2.1.2 as:

$$2 \left[ \sum_{i=1}^{n} \hat{g}(x_{i}^{(j)^T} \hat{\Sigma}_{j}^{-1} \tilde{\theta}_{j}) x_{i}^{(j)^T} \hat{\Sigma}_{j}^{-1} \hat{\Sigma}_{j,k} \hat{\Sigma}_{k}^{-1} x_{i}^{(k)} \right]$$

where in our simulation exercise $j = 3, k = 4, n = 1000$.

Out of 100 simulations, the average of the values of (9) found in the simulations is: 5.837 (0.0025). Thus roughly in line with 6.059, the mean of the s.d.o, and we also observe a limited standard deviation.

In the end we can conclude that the analytical estimator for the binomial deviance divergence optimism, in the framework outlined in Section 2.2 and to the extension of this simulation study, is slightly biased but that the approximations employed in the derivation of the formula are working relatively well.
2.2.3 Hamming Distance

In the Hamming distance case, the divergence optimism is $2 \text{tr}(\hat{\text{cov}}(\hat{y}_j, \hat{y}_k))$. Thus the simulated divergence optimism becomes:

$$2 \sum_{i=1}^{n} \hat{\text{cov}}(\hat{\pi}_i^{(j)}, \hat{\pi}_i^{(k)})$$  \hspace{1cm} (10)

where, with $l = j, k$, $\hat{\pi}_i^{(l)}$ is a binary vector whose values are determined by both $\hat{\pi}_i^{(l)}$, the vector of fitted probabilities derived from the logistic regression on model $M_l$, and the choice of the cutoff point $C_0 \in (0, 1)$. In our simulation exercise $j = 3$, $k = 4$, $n = 1000$ and $C_0 = 0.5$.

In Table 3 are provided the summary statistics over the 100 simulations of the Hamming distance simulated divergence optimism (see (10)) under the setting explained in Section 2.2.

<table>
<thead>
<tr>
<th>Optimism</th>
<th>Min</th>
<th>1st Quart</th>
<th>Median</th>
<th>Mean</th>
<th>3rd Quart</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simulated</td>
<td>51.74</td>
<td>53.35</td>
<td>54.01</td>
<td>54.04</td>
<td>54.64</td>
<td>57.02</td>
</tr>
</tbody>
</table>

Table 3: Summary statistics for the Hamming distance simulated divergence optimism (see (10)) under the simulation setting explained in Section 2.2

As done previously for the other error measures, we now compare the mean of the simulated divergence optimism (in bold in Table 3) to the value of the Hamming distance divergence optimism estimator derived in Section 2.1.3:

$$2 \sum_{i=1}^{n} \{\Phi_{d_i}[\hat{c}_{i,j}, \hat{c}_{i,j}] - \Phi_{d_i}[\hat{c}_{i,j}, \hat{c}_{i,j}]\}$$  \hspace{1cm} (11)

Out of 100 simulations, the average of the values of (11) found in the simulations is: 41.21 (2.30). Thus a value far from 54.04, the mean of the s.d.o, and we also observe a greater standard deviation than in the L2 and binomial deviance case. We conjecture that this negative result is driven by the Normal approximation (Efron, 1986 (4.16)) that has been employed in Section 2.1.3 for the closed form derivation of the divergence optimism.

In the end, we conclude that the divergence optimism estimator, even in a context where Efron’s approximation holds, is not stable and that we should find another solution for the specific case of the Hamming distance at the model selection stage. This has been done by employing directly the simulated divergence optimism (see (10)) which guarantees a good performance at the cost of a higher computational power. Besides we have to underline that this is
the only error measure for which such a procedure is computationally feasible because both $\hat{y}_{i,j}$ and $\hat{y}_{i,k}$ are indicator variables thus allowing simulations directly from the fitted probabilities $\hat{\pi}_{i,j}$ and $\hat{\pi}_{i,k}$. Unfortunately this is not the case for the other two error measures: indeed it has been of the uttermost importance to have derived working analytical formulae for both of them in Section (2.1.1) and (2.1.2).
3 Model Selection via the PDC with Oracle Ordering

In this section we will perform Model Selection in a sparse setting with the three estimators of the PDC derived in the previous paragraphs, knowing in advance the true order (i.e. oracle ordering) in which the explanatory variables will be added to the model at hand. Within this framework, we are able to divide the ordering procedure and the Model Selection phase, thus avoiding a potential source of variability in our results. Indeed doing Model Selection, as if the order was known, allow us to evaluate directly the performance of the PDC estimators in order to judge if they are valuable instruments or not.

In the next paragraph we will present the setting of the simulation study with three varying parameters, namely the sample size $n$, the number of available explanatory variables $p$ and $\rho$ the correlation between the covariates. Then, after a discussion on the PDC stopping rule, we will present the performance of the three PDC estimators for variable selection focusing on the relationship between $n$ and $p$ which is a key factor for Model Selection in high dimensions (e.g. Avella-Medina and Ronchetti, 2017).

3.1 Simulation Setting

We have taken inspiration for the simulation setting from Friedman, J., Hastie, T. and Tibshirani, R. (2010). We have generated a matrix $X_{n \times p}$ from a standard normal MVN\(^1\). We also allow for a correlation structure $\Sigma = [\sigma_{lm}]_{l,m=1, \ldots, p}$ of an autoregressive form with $\sigma_{lm} = \rho^{|l-m|}$, with $\rho = [0 0.2 0.5]$.

The responses $Y_i$ are then generated as independent Bernoulli random variables associated to the fitted probabilities obtained on the linear predictor $x_i^T \beta$ with logit link function. We choose $\beta = [2 - 1.8 1.3 0 \ldots 0]$ in order to achieve a majority of fitted values within the range $[0.2,0.8]$ as explained in Section 2.2.

We have performed simulations in two different settings: one when $n > p$ with $n = [100 \ 1000]$ and $p = [10 \ 100]$ (so that $p/n$ is fixed) and one when $n < p$ with $n = [100 \ 1000]$ and $p = [300 \ 3000]$ (again so that $p/n$ is fixed). These two settings are reproduced for each value of $\rho = [0 0.2 0.5]$.

\(^1\)However we have not employed an error term $\varepsilon_i$ for the generation of the $y$ vector in the logistic regression framework as Friedman et al. (2010) have done. This because we cannot control the impact of this error even if we generate it as a standard normal random variable since we need to pass it through the logistic function thus having to deal with this quantity $\frac{\exp(x_i^T \beta + \varepsilon_i)}{1 + \exp(x_i^T \beta + \varepsilon_i)}$ which has no known distribution to our knowledge. We have however also performed simulations with this error term and the results are consistent with the ones that we will present in this Section.
3.2 PDC Stopping Rule

Before looking at simulations results, it is worth spending some lines on the PDC stopping rule. This because the algorithm for model selection based on the PDC should, given an order of the explanatory variables that in our case is known in advance, evaluate the PDC sequence of nested models $\hat{PDC}_{1,2}$, $\hat{PDC}_{2,3}$, $\hat{PDC}_{3,4}$, ..., $\hat{PDC}_{j,k}$ (with $j < k$) and stop whenever the new evaluation of the $\hat{PDC}_{j,k}$ is increasing with respect to the previous one. However we have evidence coming from simulations that a stopping rule based on the minimization of the above sequence, achieves better results for every PDC estimators derived in this work especially in a correlated environment and when the sample size is small.

That is why we have opted for the $\min(\hat{PDC}_{1,2}, ..., \hat{PDC}_{j,k})$ stopping rule in both this section and the next one where the comparison of the PDC estimators with 10-fold CV under lasso ordering will take place.

At present time, the reasons why the $\min PDC$ performs better than the $\text{increase PDC}$ stopping rule are still under investigation. Nevertheless we can try to shed light on the matter looking at a specific simulation result coming from the L2 PDC estimator in a setting with $n = 1000$, $p = 10$, $\rho = 0$ (and $\beta = [2 -1.8 1.3 0 \ldots 0]$) meaning that we have three active regressors.\footnote{We use the term active regressor to denote the regressors with corresponding non nil slope.}

The variables are considered in their natural order, i.e. following the sequence $x_1, x_2, x_3, \ldots, x_p$. Figure 1 illustrates the evolution of the estimated PDC in this setting for one simulated sample (for the other samples, the behavior has been found to be the same). The $\hat{PDC}$ decreases until the 4th variable is added, which is the first non active regressor. Both $\min PDC$ and $\text{increase PDC}$ stopping rules are equivalent in selecting the first 3 regressors.

If correlations are introduced among (all) the covariates, the situation changes as illustrated in Figure 2 with $\rho = 0.5$. In this case, it is obvious that the correct selection criterion is when the $\hat{PDC}$ achieves its minimum value. Figure 3, illustrates the evolution of the $\hat{PDC}$ when there are more active regressors (without correlations) with $\beta = [2 -1.7 1.5 -2.5 1.5 0 \ldots 0]$, a case where the proportion of fitted probabilities in the interval $[0.2, 0.8]$ (see Section 2.2) is smaller. Again, the best selection criterion is when the $\hat{PDC}$ achieves its minimum value.

By means of these simulated samples in different conditions, we have evidence that the right criterion to prefer for model selection (i.e. when to stop) is by considering the model achieving the minimum value for the estimated PDC. This rule comes however at a computational cost since we have to evaluate the whole PDC sequence of values in order to take the minimum, which is
Figure 1: $\hat{PDC}$ for the L2 norm for one simulated sample under the simulation setting with $n = 1000$, $p = 10$, $\rho = 0$ and $\beta = [2 \ -1.8 \ 1.3 \ 0 \ldots \ 0]$. 

Figure 2: $\hat{PDC}$ for the L2 norm for one simulated sample under the simulation setting with $n = 1000$, $p = 10$, $\rho = 0.5$ and $\beta = [2 \ -1.8 \ 1.3 \ 0 \ldots \ 0]$. 

not the case when the stopping rule is increase PDC. A possible solution to avoid computing the whole sequence, comes from a common characteristic that can be easily spotted looking at Figures 1 to 3. Indeed in each sequence we
Figure 3: $\hat{PDC}$ for the L2 norm for one simulated sample under the simulation setting with $n = 1000$, $p = 10$, $\rho = 0$ and $\beta = [2 - 1.7 - 1.5 - 2.5 1.5 0 \ldots 0]$. Have evidence of a significant drop in the value of the $\hat{PDC}$ as the last active regressor is included in the model. This feature could be exploited to design an automatic procedure that would avoid to compute the estimated PDC for all the sequence of nested models. The design of this automatic procedure is left for further researches.

3.3 Simulation Results

To evaluate the performance of the selection procedures, we consider three criteria that are commonly used (e.g. Tibshirani, 1996). Namely, we consider the proportion of times the correct model is selected ($\text{Exact}$), the proportion of times the selected model contains the correct one ($\text{Correct}$) and the average number of selected regressors ($\text{Average} \#$). We compare here the three PDC estimators: the L2 norm ($\text{PDC L2}$), the Binomial deviance ($\text{PDC BIN. DEV.}$) and the simulated Hamming distance ($\text{PDC HAM. SIM.}$).

We have experienced two kinds of numerical issues both in this context and in Section (4.2): the first one is regarding the inversion of the variance matrices $\hat{\Sigma}_j$ and $\hat{\Sigma}_k$, which appear in both the L2 norm and binomial deviance PDC estimators, when the evaluated models $M_j$ and $M_k$ are high dimensional. This can be explained by the fact that at the moment, the stopping rule is the minimum of the $\hat{PDC}$ and therefore large models need to be estimated even in sparse settings (as the ones considered for the simulations). Hence, in order to have for now a workable method, we stop the search whenever a numerical
issue (warning or error) provided by R appears. This is certainly not a suitable manner to stop the search, but for the purpose of this work, it allows to study the performance of the different PDC estimators as selection criteria, and we leave aside the construction of a stopping rule for further researches.

3.3.1 When \( n > p \)

In Tables 4 and 5 are presented the performances of the three PDC estimators when \( n = 100, p = 10 \), and, respectively, \( n = 1000, p = 100 \), with \( \rho = [0 0.2 0.5] \) and \( \beta = [2 -1.8 1.3 0 \ldots 0] \).

<table>
<thead>
<tr>
<th>Divergence</th>
<th>Exact</th>
<th>Correct</th>
<th>Average ♦</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \rho = 0.2 )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PDC L2</td>
<td>0.58</td>
<td>1</td>
<td>3.86</td>
</tr>
<tr>
<td>PDC BIN. DEV.</td>
<td>0.7</td>
<td>0.85</td>
<td>3.25</td>
</tr>
<tr>
<td>PDC HAM. SIM.</td>
<td>0.11</td>
<td>0.82</td>
<td>5.31</td>
</tr>
<tr>
<td>( \rho = 0.5 )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PDC L2</td>
<td>0.62</td>
<td>1</td>
<td>3.64</td>
</tr>
<tr>
<td>PDC BIN. DEV.</td>
<td>0.59</td>
<td>0.66</td>
<td>2.44</td>
</tr>
<tr>
<td>PDC HAM. SIM.</td>
<td>0.16</td>
<td>0.74</td>
<td>4.62</td>
</tr>
</tbody>
</table>

Table 4: Performance of the PDC using respectively the L2 norm (PDC L2), the binomial deviance (PDC BIN. DEV.) and the simulated Hamming distance (PDC HAM. SIM.), under the simulation setting with \( n = 100, p = 10 \), \( \beta = [2 -1.8 1.3 0 \ldots 0] \) for \( \rho = [0 0.2 0.5] \). The performance measured are: the proportion of times the correct model is selected (Exact), the proportion of times the selected model contains the correct one (Correct) and the average number of selected regressors (Average ♦).

In the smaller sample situation (see Table 4), we have already some indication of the peculiarities of each error measure. The PDC L2 is the best in terms of correct model selection, has a high probability of finding the true model at each attempt, even in a strongly correlated environment, and exhibits a small average number of selected regressors never reaching a value of four. On the other hand, the PDC binomial deviance tends on average to select a smaller model than the PDC L2 and sometimes exhibits even an higher percentage of exact models selected. Looking at the decreasing sequence (i.e. 0.85, 0.66, 0.62) of correct models selected, we notice that the PDC binomial deviance is not robust to an increase in the correlation of the simulation environment at least compared to the PDC L2 results. Lastly the simulated PDC Hamming is inferior to the PDC L2 results for every criterion evaluated, it tends to select
larger models but it is less affected than the PDC binomial deviance from an increase in the correlation of the simulation environment.

In the larger sample setting (see Table 5), while the ratio $p/n$ is kept fixed, we observe an improved performance of the PDC L2 in every criteria evaluated especially in the percentage of exact model selected. However the PDC binomial deviance is the clear winner in this setting, at least when the correlation of the environment stays moderate. It reaches perfect results in both exact and correct model selection, it seems to have gained more than the other PDC estimators by the increased sample sizes. Nevertheless there is a great deterioration of the performance in a highly correlated environment where the PDC binomial deviance selects basically only one significant explanatory variable out of the three of the true model. This is an important drawback for real applications where the regressors are often strongly correlated but the robustness of the PDC L2 is an encouraging result in this direction. Lastly the simulated PDC Hamming remains inferior to the other two PDC estimators as already found in the previous setting. We conjecture that this is the case in general because, for the very nature of the formula, it tends to smooth the differences in the fitted probabilities. An easy example can clarify this concept: suppose that we have derived two vectors of fitted probabilities $\hat{\pi}^{(j)} = [0.4, 0.1, 0.9]$ and $\hat{\pi}^{(k)} = [0.1, 0.4, 0.6]$ for the evaluation of the two models $M_j$ and $M_k$. Then let us fix the value of $C_0 = 0.5$ and focus only on the apparent divergence part of the formula for the PDC Hamming: since we will round the values depending

---

<table>
<thead>
<tr>
<th>Divergence</th>
<th>Exact</th>
<th>Correct</th>
<th>Average $#$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho = 0$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PDC L2</td>
<td>0.77</td>
<td>1</td>
<td>3.33</td>
</tr>
<tr>
<td>PDC BIN. DEV.</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>PDC HAM. SIM.</td>
<td>0.12</td>
<td>0.87</td>
<td>5.78</td>
</tr>
<tr>
<td>$\rho = 0.2$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PDC L2</td>
<td>0.71</td>
<td>1</td>
<td>3.35</td>
</tr>
<tr>
<td>PDC BIN. DEV.</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>PDC HAM. SIM.</td>
<td>0.09</td>
<td>0.9</td>
<td>6.17</td>
</tr>
<tr>
<td>$\rho = 0.5$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PDC L2</td>
<td>0.76</td>
<td>1</td>
<td>3.28</td>
</tr>
<tr>
<td>PDC BIN. DEV.</td>
<td>0.05</td>
<td>0.05</td>
<td>1.1</td>
</tr>
<tr>
<td>PDC HAM. SIM.</td>
<td>0.17</td>
<td>0.86</td>
<td>5.28</td>
</tr>
</tbody>
</table>

Table 5: Performance of the $\hat{PDC}$ using respectively the L2 norm ($PDC\ L2$), the binomial deviance ($PDC\ BIN.\ DEV.$) and the simulated Hamming distance ($PDC\ HAM.\ SIM.$), under the simulation setting with $n = 1000$, $p = 100$, $\beta = [2 - 1.8 1.3 0 \ldots 0]$ for $\rho = [0, 0.2, 0.5]$. The performance measured are: the proportion of times the correct model is selected ($Exact$), the proportion of times the selected model contains the correct one ($Correct$) and the average number of selected regressors ($Average\ #$).
on $C_0$, the two vectors would be considered the same (i.e. $[0 0 1]$) thus giving a null apparent divergence. However in reality $\hat{\pi}^{(j)}$ and $\hat{\pi}^{(k)}$ are very different and others estimators as the PDC L2 can capture this difference.

3.3.2 When $n < p$

In Tables 6 and 7 are presented the performances of the three PDC estimators when $n = 100$, $p = 300$, and, respectively, $n = 1000$, $p = 3000$, with $\rho = [0 0.2 0.5]$ and $\beta = [2 -1.8 1.3 0 \ldots 0]$.

<table>
<thead>
<tr>
<th>Divergence</th>
<th>Exact</th>
<th>Correct</th>
<th>Average ♯</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho = 0$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PDC L2</td>
<td>0.5</td>
<td>1</td>
<td>4.64</td>
</tr>
<tr>
<td>PDC BIN. DEV.</td>
<td>0.61</td>
<td>0.7</td>
<td>2.83</td>
</tr>
<tr>
<td>PDC HAM. SIM.</td>
<td>0.1</td>
<td>0.78</td>
<td>5.36</td>
</tr>
<tr>
<td>$\rho = 0.2$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PDC L2</td>
<td>0.51</td>
<td>1</td>
<td>4.11</td>
</tr>
<tr>
<td>PDC BIN. DEV.</td>
<td>0.64</td>
<td>0.73</td>
<td>2.66</td>
</tr>
<tr>
<td>PDC HAM. SIM.</td>
<td>0.14</td>
<td>0.74</td>
<td>5.1</td>
</tr>
<tr>
<td>$\rho = 0.5$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PDC L2</td>
<td>0.6</td>
<td>0.95</td>
<td>3.56</td>
</tr>
<tr>
<td>PDC BIN. DEV.</td>
<td>0.26</td>
<td>0.31</td>
<td>1.75</td>
</tr>
<tr>
<td>PDC HAM. SIM.</td>
<td>0.1</td>
<td>0.82</td>
<td>5.61</td>
</tr>
</tbody>
</table>

Table 6: Performance of the $\hat{PDC}$ using respectively the L2 norm (PDC L2), the binomial deviance (PDC BIN. DEV.) and the simulated Hamming distance (PDC HAM. SIM.), under the simulation setting with $n = 100$, $p = 300$, $\beta = [2 -1.8 1.3 0 \ldots 0]$ for $\rho = [0 0.2 0.5]$. The performance measured are: the proportion of times the correct model is selected (Exact), the proportion of times the selected model contains the correct one (Correct) and the average number of selected regressors (Average ♯).

In the smaller sample situation (see Table 6), we observe the same features highlighted in the previous paragraph which now become specific characteristics of each error measure. The PDC L2 is still robust to correlated environments, exhibits a good result in terms of correct models which is directly related to consistency in the model selection context and, if it is not selecting the exact model, is not going too far as we can notice from the average number of regressors chosen which has been always smaller than 5. On the other hand PDC binomial deviance confirms both its tendency to deteriorate in a correlated environment and to under select in small samples. However it has an higher proportion of exact selections with respect to the PDC L2 in moderately correlated environments and this is a remarkable result. Regarding the simulated PDC Hamming, we confirm its inferiority compared to the other two PDC estimators for the same reasons highlighted in Section 3.3.1.
Table 7: Performance of the $\hat{PDC}$ using respectively the L2 norm ($PDC\ L2$), the binomial deviance ($PDC\ BIN.\ DEV.$) and the simulated Hamming distance ($PDC\ HAM.\ SIM.$), under the simulation setting with $n = 1000$, $p = 3000$, $\beta = [2 -1.8\ 1.3\ 0\ \ldots\ 0]$ for $\rho = [0\ 0.2\ 0.5]$. The performance measured are: the proportion of times the correct model is selected (Exact), the proportion of times the selected model contains the correct one (Correct) and the average number of selected regressors (Average $\#$).

<table>
<thead>
<tr>
<th>Divergence</th>
<th>Exact</th>
<th>Correct</th>
<th>Average $#$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho = 0$</td>
<td>0.63</td>
<td>1</td>
<td>3.48</td>
</tr>
<tr>
<td>$PDC\ L2$</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>$PDC\ HAM.\ SIM.$</td>
<td>0.06</td>
<td>0.91</td>
<td>6.32</td>
</tr>
<tr>
<td>$\rho = 0.2$</td>
<td>0.7</td>
<td>1</td>
<td>3.39</td>
</tr>
<tr>
<td>$PDC\ L2$</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>$PDC\ HAM.\ SIM.$</td>
<td>0.11</td>
<td>0.9</td>
<td>6.09</td>
</tr>
<tr>
<td>$\rho = 0.5$</td>
<td>0.73</td>
<td>1</td>
<td>3.33</td>
</tr>
<tr>
<td>$PDC\ L2$</td>
<td>0.02</td>
<td>0.02</td>
<td>1.04</td>
</tr>
<tr>
<td>$PDC\ HAM.\ SIM.$</td>
<td>0.1</td>
<td>0.89</td>
<td>6</td>
</tr>
</tbody>
</table>

In the larger sample setting (see Table 7), while the ratio $p/n$ is kept fixed, we observe a strong increase in the performance of the PDC L2 with respect to the small sample setting. It reaches in every case full proportion of correct models thus showing even more signs of its consistency as a model selection criterion. On the other hand PDC binomial deviance greatly benefits from the increase in the sample size demonstrating a perfect score as long as the simulated environment stays moderately correlated. The judgment on the simulated PDC Hamming remains unchanged after this round of simulations.

To conclude it is interesting to notice that all the three PDC estimators results seem not to be affected by the increase in the number of variables. This we believe shows the strength and the potential of the PDC: since it compares one by one the nested models, it does not appear to be affected by the actual number of potential covariates and the dimension of the problem is automatically reduced by the very nature of the instrument. Still we have to remember that all these results have been derived knowing in advance the true order which is not an easy matter to deal with. But after this section, we have evidence that, given the right order, the PDC L2 (and the PDC binomial deviance in moderately correlated settings) does select the exact model at least 50% of the times and it is a consistent model selection procedure with a low average of selected explanatory variables.
4 PDC and CV Comparison under Lasso Ordering

In this section, we will perform Model Selection in a sparse setting to compare the performance of the three estimators of the PDC, derived in the previous paragraphs, with 10-fold cross validation (from now on CV) which is the default criterion for model selection in the R package glmnet (Friedman, Hastie and Tibshirani, 2017). In particular we consider two different rules for the choice of the regularization parameter $\lambda$: the first one is the well known $\lambda \text{.min}$ rule. It implies to pick the value of $\lambda$ that minimizes the CV error which is the error averaged over the 10 folds. The second one is the $\lambda \text{.1SE}$ rule which requires to find the usual minimizer $\lambda \text{.min}$ and then to calculate $\lambda \text{.1SE}$ as the largest $\lambda$ such that the CV error curve is still within one standard error from $\lambda \text{.min}$. The rationale behind the $\lambda \text{.1SE}$ rule, which is preferred in sparse settings (e.g. Friedman et al., 2010), is to choose the simplest model whose accuracy is comparable with the best model. Regarding the ordering rule, the one provided by the lasso for the logistic regression (Friedman et al., 2010) of the same glmnet package is the natural choice.

In the next paragraph we will present the setting of the simulation study. It will be chosen to obtain a correct ordering (i.e. when all the active regressors are placed before the inactive ones) from the binary lasso algorithm which guarantees a meaningful comparison between CV and PDC methods. In the end we will analyze the variable selection performance of the 10-fold CV, with the two different choices for $\lambda$, and of the three estimators of the PDC focusing on the correlation present in the simulation environment.

4.1 Simulation Setting

We have reproduced the simulation setting presented in Section (3.1). We have generated a matrix $X_{n \times p}$ from a standard MVN allowing for a correlation structure $\Sigma = [\sigma_{lm}]_{l,m=1,...,p}$ of an autoregressive form with $\sigma_{lm} = \rho^{|l-m|}$, with $p = [0 0.2 0.5]$. The responses $Y_i$ are then generated as independent Bernoulli random variables associated to the true (fitted) probabilities obtained on the linear predictor $x_i^T \beta$ with logit link function. We choose $\beta = [2 - 1.8 1.3 0 \ldots 0]$ in order to achieve a majority of fitted values within the range $[0.2, 0.8]$ as explained in Section 2.2.

To ensure a meaningful comparison between CV and PDC methods, we choose a setting where $n = 1000$ and $p = 100$. This because when either the sample size $n$ is small or $p$ approaches $n$, and a fortiori when $p > n$, the ordering of the lasso algorithm deteriorates thus placing inactive regressors before the active ones. On the other hand when we opt for $(n, p) = [1000, 100]$, we are
sure to obtain at every iteration a correct ordering\textsuperscript{3}.

4.2 Simulation Results

To evaluate the performance of the selection procedures, we consider four criteria that are commonly used (e.g. Tibshirani, 1996). Namely, the proportion of times the correct model is selected (\textit{Exact}), the proportion of times the selected model contains the correct one (\textit{Correct}), the average number of selected regressors (\textit{Average \#}) and the mean square error (\textit{MSE}). We compare here 10-fold CV with the regularization parameter $\lambda$ chosen to minimize the CV error ($CV\,\lambda_{\text{min}}$), the same 10-fold CV with the regularization parameter $\lambda$ chosen to be distant one standard error from the minimizer ($CV\,\lambda_{1SE}$) and the three PDC estimators: the L2 norm ($PDC\,L2$), the Binomial deviance ($PDC\,\text{BIN. DEV.}$) and the simulated Hamming distance ($PDC\,\text{HAM. SIM.}$). In Table 8 are presented the performances of the two CV methods and of the three PDC estimators when $n = 1000$, $p = 100$, with $\rho = [0\,0.2\,0.5]$ and $\beta = [2\,-1.8\,1.3\,0\,\ldots\,0]$.

When the simulation setting is moderately correlated (i.e. $\rho = [0\,0.2]$), the PDC binomial deviance is the clear winner. When $\rho = 0$, it reaches perfect results in both exact, correct model selection and average regressors chosen together with a negligible mean squared error. Then we observe a little deterioration when $\rho = 0.2$ as already noticed in Section (3.3.1). Indeed as $\rho$ increases, PDC binomial deviance shows a tendency to under select and, as a byproduct, a larger mean squared error. The second best, as expected in sparse settings (Friedman et al., 2010), is $CV\,\lambda_{1SE}$ which exhibits a good and relatively stable performance in both exact and correct models selection. Besides, since lasso estimates are known to be far for example from the GLM ones, on which all the three PDC estimators are based, we are not surprised from the poor mean squared error performance, that it shares with $CV\,\lambda_{\text{min}}$. With respect to PDC L2, we confirm the characteristics outlined in Section (3.3.1). We observe a lower number of exact selections than in Section (3.3.1), probably related to the lasso ordering technique, a perfect selection of correct regressors and of course a low mean squared error for the same reasons stated above. Finally both $CV\,\lambda_{\text{min}}$ and PDC Hamming are inferior methods compared to the others. As expected, $CV\,\lambda_{\text{min}}$ selects way more variables than needed, thus reaching consistency but lacking in the number of exact selections. On the contrary, PDC Hamming shows the same tendency to smooth the differences outlined in Section (3.3.1).

When the simulation setting is highly correlated (i.e. $\rho = 0.5$), the PDC L2 becomes the best method. It preserves a fair number (i.e. 0.21) of exact

\textsuperscript{3}In our simulation exercise, since $\beta = [2\,-1.8\,1.3\,0\,\ldots\,0]$, a correct ordering is achieved whenever the vector which contains the sequence of variables to evaluate, starts with the triplet (1, 2, 3) where the order in which 1, 2 or 3 appear does not matter.
Table 8: Performance of the 10-fold CV with the regularization parameter $\lambda$ chosen to minimize the CV error ($CV_{\lambda \text{min}}$), of the same 10-fold CV with the regularization parameter $\lambda$ chosen to be distant one standard error from the minimizer ($CV_{\lambda \text{.1SE}}$), of the $\hat{PDC}$ using respectively the L2 norm ($PDC_{L2}$), the binomial deviance ($PDC_{BIN. DEV.}$) and the simulated Hamming distance ($PDC_{HAM. SIM.}$), under the simulation setting with $n = 1000$, $p = 100$, $\beta = [2 - 1.8 1.3 0 \ldots 0]$ for $\rho = [0 0.2 0.5]$. The performance measured are: the proportion of times the correct model is selected ($Exact$), the proportion of times the selected model contains the correct one ($Correct$), the average number of selected regressors ($Average \#$) and the mean squared error ($MSE$).

<table>
<thead>
<tr>
<th>Methods</th>
<th>Exact</th>
<th>Correct</th>
<th>Average $#$</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$CV_{\lambda \text{min}}$</td>
<td>0</td>
<td>1</td>
<td>17.27</td>
<td>7.16</td>
</tr>
<tr>
<td>$CV_{\lambda \text{.1SE}}$</td>
<td>0.63</td>
<td>1</td>
<td>3.72</td>
<td>7.34</td>
</tr>
<tr>
<td>$PDC_{L2}$</td>
<td>0.12</td>
<td>1</td>
<td>5.77</td>
<td>0.24</td>
</tr>
<tr>
<td>$PDC_{BIN. DEV.}$</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>0.07</td>
</tr>
<tr>
<td>$PDC_{HAM. SIM.}$</td>
<td>0.07</td>
<td>0.92</td>
<td>6.27</td>
<td>0.47</td>
</tr>
<tr>
<td>$CV_{\lambda \text{min}}$</td>
<td>0</td>
<td>1</td>
<td>17.68</td>
<td>7.01</td>
</tr>
<tr>
<td>$CV_{\lambda \text{.1SE}}$</td>
<td>0.69</td>
<td>1</td>
<td>3.49</td>
<td>7.23</td>
</tr>
<tr>
<td>$PDC_{L2}$</td>
<td>0.21</td>
<td>1</td>
<td>5.24</td>
<td>0.16</td>
</tr>
<tr>
<td>$PDC_{BIN. DEV.}$</td>
<td>0.92</td>
<td>0.92</td>
<td>2.84</td>
<td>0.5</td>
</tr>
<tr>
<td>$PDC_{HAM. SIM.}$</td>
<td>0.1</td>
<td>0.9</td>
<td>6.18</td>
<td>0.72</td>
</tr>
<tr>
<td>$CV_{\lambda \text{min}}$</td>
<td>0</td>
<td>1</td>
<td>23.73</td>
<td>6.76</td>
</tr>
<tr>
<td>$CV_{\lambda \text{.1SE}}$</td>
<td>0.11</td>
<td>1</td>
<td>6.22</td>
<td>7.05</td>
</tr>
<tr>
<td>$PDC_{L2}$</td>
<td>0.21</td>
<td>1</td>
<td>4.85</td>
<td>0.14</td>
</tr>
<tr>
<td>$PDC_{BIN. DEV.}$</td>
<td>0.01</td>
<td>0.01</td>
<td>1.02</td>
<td>5.85</td>
</tr>
<tr>
<td>$PDC_{HAM. SIM.}$</td>
<td>0.1</td>
<td>0.86</td>
<td>5.76</td>
<td>0.88</td>
</tr>
</tbody>
</table>

selections, shows consistency (i.e. perfect selection of correct regressors) and a low number of average selected regressors compared to the other methods keeping out the PDC binomial deviance case. Indeed the latter, shows the same feature outlined in Section (3.3.1): it is not robust to an increase in the correlation of the simulation environment, at least compared to the results of the other methods. The second best is again $CV_{\lambda \text{.1SE}}$ whose performance in the exact models selected deteriorates a lot when $\rho = 0.5$. In the end, this increase in the correlation, has not modified our judgment on both $CV_{\lambda \text{min}}$ and PDC Hamming. However PDC Hamming, although still inferior to PDC L2, is at least showing some robustness to this correlated setting.

To conclude, we have evidence from Table 8, and up to the extension of this simulation study, that, given a correct order, the PDC binomial deviance in
moderately correlated settings and the PDC L2 in highly correlated ones, exhibit a better model selection performance than both CV $\lambda_{\text{min}}$ and CV $\lambda_{1SE}$ which are the default methods in the R package glmnet (Friedman et al., 2017). Moreover the robustness of the PDC L2 performance, observed both in Table 8 and in Sections (3.3) in highly correlated settings, is a promising sign for real applications where these scenarios are often the norm and not the exception.
5 Conclusion

In this Master Thesis, we have analytically derived and numerically implemented three estimators of the Prediction Divergence Criterion (Avella-Medina et al., working paper) for Model Selection within the logistic regression framework. After the validation of these estimators by means of simulations, we have performed Model Selection both when the order of the variables was known in advance and when the order was correct but decided by an already existing algorithm, namely the binary lasso (Friedman et al., 2010). Finally we have produced evidences of the good performance of two of these estimators, one derived from the L2 norm error measure and the other from the binomial deviance, respectively in highly and moderately correlated settings. They have been proven better, to the extension of the simulation study of Section (4), than the defaults methods, based on 10-fold Cross Validation, currently available in the glmmnet (Friedman et al., 2017) R package.

Among the limitations of this work, we can consider both the lack of an implementation of a reliable automatic procedure for the stopping of each of the three estimated Prediction Divergence Criteria and the limited extension of the simulations studies done in Sections (3) and (4). Besides, we have not dealt with the problem of selecting a specific ordering method, leaving it for further researches.

Regarding the future research paths, we would like to perform Model Selection with the three derived PDC criteria in a real dataset. Furthermore we are considering to search an approximation of the second order (e.g. a 2nd order Taylor expansion) to improve the quality of the estimated quantities. Finally, since Efron’s approximation (1978, eq. (2.4)) is applicable to the whole domain of exponential families, it may be interesting to derive an estimator of the Prediction Divergence Criterion for the Generalized Linear Model case or more modestly to extend this work to the Poisson regression framework.
References


32


A Analytical Derivations

A.1 L2 Norm

It is worth starting from equation (2) for the L2 case:
\[
PDC_{j,k} = ||\hat{y}_j - \hat{y}_k||_2^2 + 2 \text{tr} [\text{cov}(\hat{y}_j, \hat{y}_k)]
\]

For the derivation of \(\text{cov}(\hat{y}_j, \hat{y}_k)\), we proceed element by element (omitting the index \(i\)):
\[
\text{cov}(\hat{y}_j, \hat{y}_k) = \mathbb{E}(\hat{y}_j \hat{y}_k) - \mathbb{E}(\hat{y}_j) \mathbb{E}(\hat{y}_k)
\]

We have for \(l = j, k\)
\[
\mathbb{E}(\hat{y}_l) = \mathbb{E}\left[\frac{\exp(x_l^T \hat{\beta}_l)}{1 + \exp(x_l^T \hat{\beta}_l)}\right]
\]

Efron, in his 1978 paper (eq. 2.4), proposes the approximation \(\hat{\beta}_l \approx \Sigma^{-1}_l \hat{\theta}_l\)

where \(\hat{\theta}_l = \sum_{i=1}^{n} x_i^T y_i\) and \(\Sigma_l = \sum_{i=1}^{n} \text{var}[y_i] x_i^T x_i^T\) with \(\text{var}[y_i] = \pi_i (1 - \pi_i).\)

We then write
\[
\mathbb{E}(\hat{y}_l) = \mathbb{E}\left[\frac{\exp(x_l^T \Sigma^{-1}_l \hat{\theta}_l)}{1 + \exp(x_l^T \Sigma^{-1}_l \hat{\theta}_l)}\right] = \mathbb{E}\left[\frac{\exp(c_l^T \hat{\theta}_l)}{1 + \exp(c_l^T \hat{\theta}_l)}\right] = \mathbb{E}[g(c_l^T \hat{\theta}_l)]
\]

A 1st order Taylor expansion of \(g(c_l^T \hat{\theta}_l)\) around \(\mathbb{E}[\hat{\theta}_l]\) leads to
\[
g(c_l^T \hat{\theta}_l) = g(c_l^T \mathbb{E}[\hat{\theta}_l]) + \hat{g}(c_l^T \mathbb{E}[\hat{\theta}_l]) c_l^T (\hat{\theta}_l - \mathbb{E}[\hat{\theta}_l]) + o (|c_l^T (\hat{\theta}_l - \mathbb{E}[\hat{\theta}_l])|)(15)
\]
so that
\[
\mathbb{E}[g(c_l^T \hat{\theta}_l)] \approx g(c_l^T \mathbb{E}[\hat{\theta}_l])
\]

with \(\mathbb{E}[\hat{\theta}_l] = \sum_{i=1}^{n} x_i^T \pi_i\). Making use of (15) for both \(\hat{y}_j\) and \(\hat{y}_k\), we also have:
\[
\mathbb{E}[\hat{y}_j \hat{y}_k] \approx \mathbb{E}\left[\left(g(c_j^T \mathbb{E}[\hat{\theta}_j]) + \hat{g}(c_j^T \mathbb{E}[\hat{\theta}_j]) c_j^T (\hat{\theta}_j - \mathbb{E}[\hat{\theta}_j])\right)\left(g(c_k^T \mathbb{E}[\hat{\theta}_k]) + \hat{g}(c_k^T \mathbb{E}[\hat{\theta}_k]) c_k^T (\hat{\theta}_k - \mathbb{E}[\hat{\theta}_k])\right)\right]
\]
\[
\approx g(c_j^T \mathbb{E}[\hat{\theta}_j]) g(c_k^T \mathbb{E}[\hat{\theta}_k]) + \hat{g}(c_j^T \mathbb{E}[\hat{\theta}_j]) \hat{g}(c_k^T \mathbb{E}[\hat{\theta}_k]) c_{jk}^T + \mathbb{E}\left[(\hat{\theta}_j - \mathbb{E}[\hat{\theta}_j]) (\hat{\theta}_k - \mathbb{E}[\hat{\theta}_k])^T\right] c_{jk}
\]

34
with

\[
\mathbb{E} \left[ (\hat{\theta}_j - \mathbb{E}[\hat{\theta}_j])(\hat{\theta}_k - \mathbb{E}[\hat{\theta}_k])^T \right] = \sum_{i=1}^{n} x_i^{(j)} (y_i - \pi_i) \sum_{i=1}^{n} x_i^{(k)} (y_i - \pi_i) \\
= \sum_{i=1}^{n} x_i^{(j)} x_i^{(k)} (y_i - \pi_i)^2 \\
= \sum_{i=1}^{n} x_i^{(j)} x_i^{(k)} T \pi_i (1 - \pi_i)
\]

(18)

Now we turn to (12) to get

\[
cov(\hat{y}_j, \hat{y}_k) \approx \hat{g} \left( c^{(j)T} \sum_{i=1}^{n} x_i^{(j)} \pi_i \right) \hat{g} \left( c^{(k)T} \sum_{i=1}^{n} x_i^{(k)} \pi_i \right) c^{(j)T} \left( \sum_{i=1}^{n} x_i^{(j)} x_i^{(k)} T \pi_i (1 - \pi_i) \right) c^{(k)}
\]

(19)

with

\[
\hat{g}(\lambda) = \frac{\exp(\lambda)}{1 + \exp(\lambda)} \left( 1 - \frac{\exp(\lambda)}{1 + \exp(\lambda)} \right)
\]

(20)

In order to use (19) for finding \( \overline{PDC}_{j,k} \), it is necessary to choose an estimator for \( \pi_i \). In line with the derivations presented in Avella-Medina et al. (working paper), we assume that the smaller model \( M_j \) is the correct one which implies \( \pi_i = \pi_i^{(j)} \). For this reason we consistently use \( \tilde{\pi}_i^{(j)} \) when \( \tilde{\pi}_i \) appears in the above formula.
A.2 Binomial Deviance

The starting point is equation (4):

\[
tr[\hat{\text{cov}}(\hat{y}_j, \nabla \psi(\hat{y}_k))] = 2 \sum_{i=1}^{n} \hat{\text{cov}}(\hat{y}_{i,j}, \log(\hat{y}_{i,k} / (1 - \hat{y}_{i,k}))
\]

For the derivation of \(\hat{\text{cov}}(\hat{y}_{i,j}, \log(\hat{y}_{i,k} / (1 - \hat{y}_{i,k}))\), we proceed element by element (omitting the index \(i\)):

\[
\text{cov}(\hat{y}_j, \log(\hat{y}_k / (1 - \hat{y}_k))) = \mathbb{E}(\hat{y}_j \log(\hat{y}_k / (1 - \hat{y}_k))) - \mathbb{E}(\hat{y}_j) \mathbb{E}(\log(\hat{y}_k / (1 - \hat{y}_k)))
\]

We make use of (13), (14), (15), (16) so that we have for \(j\):

\[
\mathbb{E}(\hat{y}_j) \approx g(c^{(j)T} \mathbb{E}[\hat{\theta}_j])
\]

with \(\mathbb{E}[\hat{\theta}_j] = \sum_{i=1}^{n} x_i^{(j)} \pi_i\).

On the other hand for \(k\), thanks to the presence of the canonical link for the logistic regression and to Efron’s (1978, eq. 2.4) approximation, we have

\[
\mathbb{E}(\log(\hat{y}_k / (1 - \hat{y}_k))) = \mathbb{E}[x^{(k)T} \hat{\beta}_k] \\
\approx x^{(k)T} \mathbb{E}[\Sigma_k^{-1} \hat{\theta}_k] \\
\approx x^{(k)T} \Sigma_k^{-1} \theta_k
\]

(21)

where \(\theta_k = \sum_{i=1}^{n} x_i^{(k)} \pi_i\). Now if we make use of equation (15) and (21) together with \(c^{(l)T} = x^{(l)T} \Sigma_l^{-1}\) for \(l = j, k\) we can write:

\[
\mathbb{E}(\hat{y}_j \log(\hat{y}_k / (1 - \hat{y}_k))) \approx \mathbb{E}[(g(c^{(j)T} \mathbb{E}[\hat{\theta}_j]) + \hat{g}(c^{(j)T} \mathbb{E}[\hat{\theta}_j]) c^{(j)T} (\hat{\theta}_j - \mathbb{E}[\hat{\theta}_j])) c^{(k)T} \hat{\theta}_k] \\
\approx (g(c^{(j)T} \mathbb{E}[\hat{\theta}_j]) c^{(k)T} \theta_k + \hat{g}(c^{(j)T} \mathbb{E}[\hat{\theta}_j]) c^{(j)T} \mathbb{E}[(\hat{\theta}_j - \mathbb{E}[\hat{\theta}_j]) \hat{\theta}_k^T]) c^{(k)}
\]
Now knowing that:

\[
\mathbb{E}[(\hat{\theta}_j - \mathbb{E}[\hat{\theta}_j]) \hat{\theta}_k^T] = \mathbb{E} \left[ \sum_{i=1}^{n} x_i^{(j)} (y_i - \pi_i) \sum_{i=1}^{n} x_i^{(k)^T} y_i \right] \\
= \mathbb{E} \left[ \sum_{i=1}^{n} x_i^{(j)} x_i^{(k)^T} (y_i^2 - \pi_i y_i) \right] \\
= \sum_{i=1}^{n} x_i^{(j)} x_i^{(k)^T} \pi_i (1 - \pi_i) \quad (22)
\]

Finally we can derive the expression:

\[
cov\left(\hat{y}_{i,j}, \log \left( \frac{\hat{y}_{i,k}}{1 - \hat{y}_{i,k}} \right) \right) \approx \hat{g} \left( c^{(j)^T} \sum_{i=1}^{n} x_i^{(j)} \pi_i \right) \\
= c^{(j)^T} \left( \sum_{i=1}^{n} x_i^{(j)^T} x_i^{(k)^T} \pi_i (1 - \pi_i) \right) c^{(k)} \quad (23)
\]

where

\[
\hat{g}(\lambda) = \frac{\exp(\lambda)}{1 + \exp(\lambda)} \left( 1 - \frac{\exp(\lambda)}{1 + \exp(\lambda)} \right)
\]

In order to use (23) for finding $\hat{PDC}_{j,k}$, it is necessary to choose an estimator for $\pi_i$. In line with the derivations presented in Avella-Medina et al. (working paper), we assume that the smaller model $M_j$ is the correct one which implies $\pi_i = \pi_i^{(j)}$. For this reason we consistently use $\hat{\pi}_i^{(j)}$ when $\hat{\pi}_i$ appears in the above formula.
A.3 Hamming Distance

It is worth starting from equation (2) for the Hamming distance case:

\[
\hat{PDC}_{j,k} = n^{-1} \sum_{i=1}^{n} \{ \hat{y}_{i,j} \neq \hat{y}_{i,k} \} + 2 \hat{\text{cov}}(\hat{y}_j, \hat{y}_k)
\]

For the derivation of \( \hat{\text{cov}}(\hat{y}_j, \hat{y}_k) \), we proceed element by element:

\[
\text{cov}(\hat{y}_{i,j}, \hat{y}_{i,k}) = \mathbb{E}[\hat{y}_{i,j} \hat{y}_{i,k}] - \mathbb{E}[\hat{y}_{i,j}] \mathbb{E}[\hat{y}_{i,k}]
\] (24)

The main difference, with respect to the L2 and binomial deviance derivations, is that \( \hat{y}_{i,j} \) or \( \hat{y}_{i,k} \) are now indicator variables that depend on the chosen cutoff point \( C_0 \in (0, 1) \) \((C_0 = 0.5 \text{ is common})\).

We use the fact that for \( l = j, k \):

\[
\hat{y}_{i,l} > C_0 = \frac{\exp(x_i^{(l)T} \hat{\beta}_l)}{1 + \exp(x_i^{(l)T} \hat{\beta}_l)} > C_0
\]

\[
= \left( \frac{C_0}{1 - C_0} \right) - x_i^{(l)T} \beta_l
\]

\[
= \left( \frac{C_0}{1 - C_0} \right) - x_i^{(l)T} \beta_l
\]

\[
= \left( \frac{C_0}{1 - C_0} \right) - x_i^{(l)T} \beta_l > c_{i,l}
\] (25)

Now thanks to (25) and the fact that \( \hat{y}_{i,l} \) is an indicator variable, we find that:

\[
\mathbb{E}[I(\hat{y}_{i,l})] = \int_{-\infty}^{+\infty} I(\hat{y}_{i,l} > C_0) f(\hat{y}_{i,l}) \, d\hat{y}_{i,l}
\]

\[
= \mathbb{P}_\pi[\hat{y}_{i,l} > C_0]
\]

\[
= \mathbb{P}_\pi[x_i^{(l)T} (\hat{\beta}_l - \beta_l) > c_{i,l}]
\] (26)

At this point we make use of Efron approximation (1978, eq. 2.4) and we can substitute \( \hat{\beta}_l \approx \Sigma_l^{-1} \hat{\theta}_l \) and \( \beta_l \approx \Sigma_l^{-1} \theta_l \) in equation (26) knowing that

\[
\hat{\theta}_l = \sum_{i=1}^{n} x_i^{(l)T} y_i \text{ and } \Sigma_l = \sum_{i=1}^{n} \text{var}[y_i] x_i^{(l)} x_i^{(l)T} \text{ with } \text{var}[y_i] = \pi_i(1 - \pi_i).
\]
Furthermore we make use of Lyapunov central limit theorem (e.g. Efron 1986, (4.16)) to find that \((\hat{\theta}_l - \theta_l) \longrightarrow \mathcal{N}(0, \Sigma_l)\) which implies that:

\[
x_i^{(l)T} \Sigma_{l}^{-1}(\hat{\theta}_l - \theta_l) \longrightarrow \mathcal{N}(0, d_l) = \sqrt{d_l} Z_l
\]

Where \(Z_l \sim \mathcal{N}(0,1)\) and \(d_l = x_i^{(l)T} \Sigma_{l}^{-1} x_i^{(l)}\).

Now we can restart from equation (26) and thanks to Efron’s result combined with (27) derive the final expression for the marginal expectation as:

\[
E_\pi [\hat{y}_{i,l}] = \mathbb{P}_\pi \left[ x_i^{(l)T} (\hat{\beta}_l - \beta_l) > c_{l,i} \right] \\
\approx \mathbb{P}_\pi \left[ x_i^{(l)T} \Sigma_{l}^{-1}(\hat{\theta}_l - \theta_l) > c_{l,i} \right] \\
\approx \mathbb{P}_\pi \left[ \sqrt{d_l} Z_l > c_{l,i} \right] \\
\approx 1 - \Phi \left( \frac{c_{l,i}}{\sqrt{d_l}} \right)
\]

Now we turn to the joint expectation of equation (24) and we use the same path followed above for the marginal expectation to find that:

\[
E_\pi [\hat{y}_{i,j} \hat{y}_{i,k}] = \mathbb{P}_\pi [\hat{y}^{(j)}_{i} > C_0, \hat{y}^{(k)}_{i} > C_0] \\
= \mathbb{P}_\pi \left[ x_i^{(j)T} (\hat{\beta}_j - \beta_j) > c_{j,i}, x_i^{(k)T} (\hat{\beta}_k - \beta_k) > c_{k,i} \right] \\
\approx \mathbb{P}_\pi \left[ x_i^{(j)T} \Sigma_{j}^{-1}(\hat{\theta}_j - \theta_j) > c_{j,i}, x_i^{(k)T} \Sigma_{k}^{-1}(\hat{\theta}_k - \theta_k) > c_{j,k} \right] \\
\approx \mathbb{P}_\pi \left[ W_{i,j} > c_{i,j}, W_{i,k} > c_{i,k} \right]
\]

where \(W_{j,i} \sim \mathcal{N}(0, d_{ij})\) and \(W_{k,i} \sim \mathcal{N}(0, d_{ik})\) are not independent.

At this point, since each marginal distribution is gaussian, the bivariate distribution is also gaussian and it is characterized by the two first moments.
We first start deriving the generic \( \text{var}_\pi(W_{l,i}) \) where \( l = \{j,k\} \):

\[
\text{var}_\pi(W_{l,i}) = \text{var}_\pi \left( x_i^{(l)} \Sigma_{l}^{-1} (\hat{\theta}_l - \theta_l) \right) \\
= x_i^{(l)} \Sigma_{l}^{-1} \text{var}_\pi \left( \hat{\theta}_l - \theta_l \right) \Sigma_{l}^{-1} x_i^{(l)} \\
= x_i^{(l)} \Sigma_{l}^{-1} \text{var}_\pi \left( \sum_{i=1}^{n} x_i^{(l)} y_i - \sum_{i=1}^{n} x_i^{(l)} \pi_i \right) \Sigma_{l}^{-1} x_i^{(l)} \\
= x_i^{(l)} \Sigma_{l}^{-1} \text{var}_\pi \left( \sum_{i=1}^{n} x_i^{(l)} y_i \right) \Sigma_{l}^{-1} x_i^{(l)} \\
= x_i^{(l)} \Sigma_{l}^{-1} \Sigma_{l} \Sigma_{k}^{-1} x_i^{(l)} \\
= x_i^{(l)} \Sigma_{l}^{-1} x_i^{(l)} \\
= d_{il}.
\]

Similarly for the covariance we have:

\[
\text{cov}_\pi(W_{i,j}, W_{i,k}) = \text{cov}_\pi \left( x_i^{(j)} \Sigma_{j}^{-1} (\hat{\theta}_j - \theta_j) x_i^{(k)} \Sigma_{k}^{-1} (\hat{\theta}_k - \theta_k) \right) \\
= x_i^{(j)} \Sigma_{j}^{-1} \text{cov}_\pi \left( \sum_{h=1}^{n} x_h^{(j)} y_h \sum_{i=1}^{n} x_h^{(k)} y_i \right) \Sigma_{k}^{-1} x_i^{(k)} \\
= x_i^{(j)} \Sigma_{j}^{-1} \left( \sum_{h=1}^{n} x_h^{(j)} \text{var}_\pi(y_h) x_i^{(k)} + \sum_{h \neq i} x_h^{(j)} \text{cov}_\pi(y_h, y_i) x_i^{(k)} \right) \Sigma_{k}^{-1} x_i^{(k)} \\
= x_i^{(j)} \Sigma_{j}^{-1} \left( \sum_{i=1}^{n} x_i^{(j)} \pi_i (1 - \pi_i) \right) \Sigma_{k}^{-1} x_i^{(k)} \\
= x_i^{(j)} \Sigma_{j}^{-1} \Sigma_{jk} \Sigma_{k}^{-1} x_i^{(k)} \\
= d_{ijk}
\]

For symmetry we have also that \( d_{ijk} = d_{ikj} \).
Now we can derive a closed form expression for equation (24). The following property is useful for the next derivations:

\[ \tilde{F}(x_1, x_2) = 1 - F_1(x_1) - F_2(x_2) + F(x_1, x_2), \]

where \( F(x_1, x_2) = \mathbb{P}(X_1 < x_1, X_2 < x_2) \) and \( \tilde{F}(x_1, x_2) = \mathbb{P}(X_1 > x_1, X_2 > x_2) \) is the bivariate survival CDF.

At this point we can work directly at the divergence optimism level (i.e. 2 \( \text{cov} [\hat{y}_{i,j}, \hat{y}_{i,k}] \)) to find:

\[
2 \text{cov}_\pi [\hat{y}_{i,j}, \hat{y}_{i,k}] \approx 2 \left( \mathbb{P}_\pi \left[ W_{j,i} > c_{j,i}, W_{k,i} > c_{k,i} \right] \right) - 2 \left( 1 - \Phi \left( \frac{c_{j,i}}{\sqrt{d_{ij}}} \right) \right) \left( 1 - \Phi \left( \frac{c_{k,i}}{\sqrt{d_{ik}}} \right) \right)
\]

\[
\approx 2 \left( 1 - \Phi \left( \frac{c_{j,i}}{\sqrt{d_{ij}}} \right) \right) - \Phi \left( \frac{c_{k,i}}{\sqrt{d_{ik}}} \right) \mathbb{P}_\pi \left[ W_{j,i} < c_{j,i}, W_{k,i} < c_{k,i} \right]
\]

\[
\approx 2 \left( 1 - \Phi \left( \frac{c_{j,i}}{\sqrt{d_{ij}}} \right) \right) - \Phi \left( \frac{c_{k,i}}{\sqrt{d_{ik}}} \right) \Phi \left( \frac{c_{j,i}}{\sqrt{d_{ij}}} \right) - \Phi \left( \frac{c_{k,i}}{\sqrt{d_{ik}}} \right) \Phi \left( \frac{c_{j,i}}{\sqrt{d_{ij}}} \right)
\]

\[
\approx 2 \Phi d_i [c_{j,i}, c_{k,i}] - 2 \Phi \left( \frac{c_{k,i}}{\sqrt{d_{ik}}} \right) \Phi \left( \frac{c_{j,i}}{\sqrt{d_{ij}}} \right)
\]

where

\[
d_i = \begin{bmatrix} d_{ij} & d_{ijk} \\ d_{ikj} & d_{ik} \end{bmatrix}
\]

and \( \Phi(a) = \int_{-\infty}^{a} (2\pi)^{-1/2} \exp(-\frac{1}{2}z^2)dz \) and \( \Phi_d(a, b) = \int_{-\infty}^{a} \int_{-\infty}^{b} (2\pi)^{-1/2}|d|^{-1/2} \exp(-\frac{1}{2}z^T d^{-1} z)dz \) with \( z \) bivariate and non-standardized.

Finally we can write the univariate theoretical divergence optimism as

\[
2 \Phi d_i [c_{j,i}, c_{k,i}] - 2 \Phi d \approx 2 \Phi d_i [c_{j,i}, c_{k,i}]
\]

where

\[
d_i = \begin{bmatrix} d_{ij} & 0 \\ 0 & d_{ik} \end{bmatrix}
\]

41
In order to use (29) for finding $\hat{PDC}_{j,k}$, it is necessary to make a choice on the estimator of $\pi_i$. In line with the derivations presented in Avella-Medina et al. (working paper), we assume that the smaller model $M_j$ is the correct one which implies $\pi_i = \pi_i^{(j)}$. For this reason we consistently use $\pi_i^{(j)}$ when $\hat{\pi}_i$ appears in the above formula. Moreover since $X_k\beta_k = X_j\beta_j$ under this hypothesis, we obtain a simplified version of (29) where $c_{j,i} = c_{k,i}$. 