A predictive deviance criterion for selecting a generative model in semi-supervised classification

Vincent Vandewalle, Christophe Biernacki, Gilles Celeux, Gérard Govaert

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

A predictive deviance criterion for selecting a generative model in semi-supervised classification

Vincent Vandewalle — Christophe Biernacki — Gilles Celeux — Gérard Govaert

N° 7377

Septembre 2010

Thème COG

ISSN 0249-6399 ISRN INRIA/RR--7377--FR+ENG

INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE

INRIA

centre de recherche
SACLAY - ÎLE-DE-FRANCE

A predictive deviance criterion for selecting a generative model in semi-supervised classification

Vincent Vandewalle* , Christophe Biernacki†, Gilles Celeux‡ , Gérard Govaert§

Thème COG — Systèmes cognitifs
Équipes-Projets SELECT

Rapport de recherche n° 7377 — Septembre 2010 — 24 pages

Abstract: Semi-supervised classification can be hoped to improve generative classifiers by taking profit of the information provided by the unlabeled data points, especially when there are far more unlabeled data than labeled data. This paper is concerned with selecting a generative classification model from both unlabeled and labeled data. We propose a predictive deviance criterion AIC_{cond} aiming to select a parsimonious and relevant generative classifier in the semi-supervised context. Contrary to standard information criteria as AIC and BIC, AIC_{cond} is focusing to the classification task since it aims to measure the predictive power of a generative model by approximating its predictive deviance. On an other hand, it avoids the computational trouble encountered with cross validation criteria due to the repeated use of the EM algorithm. AIC_{cond} is proved to have consistency properties ensuring its parsimony compared to the Bayesian Entropy Criterion (BEC) which has a similar focus than AIC_{cond}. In addition, numerical experiments on both simulated and real data sets highlight an encouraging behavior of AIC_{cond} for variable and model selection in comparison to the other mentioned criteria.

Key-words: Generative Models, Gaussian Mixture Models, Maximum Likelihood, EM Algorithm, Cross-Validated Error Rate, Information Criteria, AIC, BIC, BEC

* CNRS & Université de Lille 1, Villeneuve d'Ascq, Université de Lille 2
† CNRS & Université de Lille 1, Villeneuve d'Ascq
‡ Inria Saclay Île-de-France
§ Université de Technologie de Compiègne

Centre de recherche INRIA Saclay – Île-de-France
Parc Orsay Université
4, rue Jacques Monod, 91893 ORSAY Cedex
Téléphone : +33 1 72 92 59 00

Un critère de déviance prédictive pour la sélection d'un modèle génératif en classification semi-supervisée

Résumé : La classification semi-supervisée donne l'opportunité d'améliorer les classifieurs génératifs par la prise en compte de l'information des points non étiquetés lorsque ceux-ci sont beaucoup plus nombreux que les points étiquetés. Cet article a trait à la sélection d'un modèle de classification génératif dans un contexte semi-supervisé. Nous proposons un critère de déviance prédictive AIC_{cond} pour choisir un modèle génératif parcimonieux de classification. Au contraire des critères classiques d'information comme AIC ou BIC, AIC_{cond} se focalise sur le but de classification en mesurant le pouvoir prédictif d'un modèle génératif par sa déviance prédictive. Par ailleurs, il évite les problèmes de temps de calcul inhérents à la validation croisée à cause de l'emploi répété de l'algorithme EM. Nous prouvons des propriétés de convergence du critère AIC_{cond} qui assurent sa supériorité vis-à-vis du critère d'entropie bayésienne BEC dont le but est analogue. De plus, des illustrations numériques sur des données réelles et simulées mettent en lumière un comportement prometteur de AIC_{cond} par rapport aux critères mentionnés pour la sélection de variables et de modèles génératifs de classification à partir d'échantillons semi-supervisés.

Mots-clés : modèles génératifs, mélanges gaussiens, maximum de vraisemblance, algorithme EM, erreur de classement évalué par validation croisée, critères d'information AIC, BIC, BEC

1 Introduction

Discriminant analysis is designing and assessing classifiers from a training set of labeled data [13] or [12]. But, more and more data sets contain in addition numerous *unlabeled data* for free. Semi-supervised classification aims to improve the classifiers performance by using the information arising from the unlabeled data and it is now an important issue in machine learning [6].

In classification, the predictive and generative approaches are in competition. The predictive approach is modelling the conditional distribution $p(z|\boldsymbol{x})$ of a class label z knowing the vector of predictors $\boldsymbol{x}$. Logistic regression [2] is an example of a semi-parametric model obtained from this predictive view point and leading to linear decision boundaries between the classes. In a similar spirit, without modelling the conditional distribution $p(z|\boldsymbol{x})$, Rosenblat Perceptron [17] and Support Vectors Machines [22] are aiming to find optimal linear decision boundaries between the classes. In one hand, by avoiding unrealistic assumptions on the joint distribution $p(\boldsymbol{x}, z)$, predictive methods could be expected to provide better classifiers in many practical situations [22]. On the other hand, they are unable to take into consideration unlabeled data without additional assumptions on the marginal distribution $p(\boldsymbol{x})$ of the predictors.

The generative approach models the joint distribution $p(\boldsymbol{x}, z)$. Examples of generative methods are Linear Discriminant Analysis [8] (LDA) and Quadratic Discriminant Analysis (QDA) which both are assuming that the class-conditional densities are Gaussian, while much more flexible models exist, see for instance [11]. In a semi-supervised setting, the generative point of view is quite natural since it leads to model the marginal distribution $p(\boldsymbol{x})$ with the mixture model $p(\boldsymbol{x}) = \sum_z p(\boldsymbol{x}, z)$ and the joint distribution form is specified by the generative model. Thus, the maximum likelihood estimate of the model parameters can be derived through the EM algorithm [14]. Moreover, when the assumptions associated to the generative model are verified, it outperforms predictive models since the whole available information is used [16]. Thus, an important question is to take profit of the unlabeled data to choose a relevant generative model.

In this purpose, many well-known model selection criteria are available, as the Bayesian Information Criterion (BIC) [18], the Akaike criterion (AIC) [1] and the V-fold cross-validated error rate [19], or the Bayesian Entropy Criterion (BEC) [4], a criterion specific to the classification context. However, BIC and AIC are known to have a possible disappointing behavior in a discriminant analysis context whereas BEC could tend to select too complex models [4]. On the contrary, the V-fold cross-validated error rate could be expected to lead to better results despite the choice of V could be sensitive and it is time consuming, especially when an iterative algorithm as EM is involved in the estimation process.

In this paper, we propose an alternative criterion aiming to estimate the predictive ability of a generative model of classification. This criterion, called AIC_{cond}, is an asymptotic AIC-like approximation of the predictive deviance of the generative model. It involves a penalty which depends on both the overlapping between the classes and the number of "predictive" parameters of the generative model. Its computational cost is similar to this one of AIC or BIC and it enjoys good consistency theoretical properties.

The paper is organized as follows. In Section 2 the generative framework is presented and standard model selection criteria are reviewed. The new criterion

AIC_{cond} is presented and theoretically studied in Section 3. Sections 4 and 5 are devoted to the presentation of numerical experiments comparing AIC_{cond} with the other criteria on simulated and real data sets, respectively. A short discussion ends the paper. Appendices contain the proofs of the propositions of Section 3.

2 Generative models in semi-supervised classification

In this section, estimation and selection of generative models are sketched in the semi-supervised classification setting.

2.1 Notation and parameter estimation

A classification problem with g classes is considered, $\mathcal{X}$ is denoting the predictors space, $\mathcal{Z} = \{1, \ldots, g\}$ the label space and $(\boldsymbol{x}, z)$ a couple of random variables on $\mathcal{X} \times \mathcal{Z}$ with probability density function (pdf) p according to some measure on $\mathcal{X} \times \mathcal{Z}$. The labeled sample of n_ℓ independent and identically distributed (iid) data arising from $(\boldsymbol{x}, z)$ is noted $\mathcal{D}_\ell = \{(\boldsymbol{x}_1, z_1), \ldots, (\boldsymbol{x}_{n_\ell}, z_{n_\ell})\}$ and the unlabeled sample of n_u iid unlabeled data arising from $\boldsymbol{x}$ is noted $\mathcal{D}_u = \{\boldsymbol{x}_{n_\ell+1}, \ldots, \boldsymbol{x}_{n_\ell+n_u}\}$. Moreover, it is important to stress that the $\boldsymbol{x}_i$'s, with $1 \leq i \leq n_\ell + n_u$, arise from the same mixture distribution. The total number of data is noted $n = n_\ell + n_u$, and the whole training sample is noted $\mathcal{D} = \{\mathbf{x}, \mathbf{z}\}$ with $\mathbf{x} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ and $\mathbf{z} = \{z_1, \ldots, z_{n_\ell}\}$. In addition, it is assumed the proportion $\beta = n_\ell/n$ of labeled sample is fixed and not depending of n. This assumption is made to give sense to asymptotic calculations in Section 3.

We consider generative parametric models $p(\boldsymbol{x}, z; \theta) = \pi_z p(\boldsymbol{x}; \eta_z)$ where π_z is the unconditional probability of the class z ($\sum_k \pi_k = 1$, $\pi_k > 0$), $p(\boldsymbol{x}; \eta_z)$ is the class-conditional pdf of $\boldsymbol{x}|z$ and $\theta = (\pi_1, \ldots, \pi_{g-1}, \eta_1, \ldots, \eta_g)$ is the whole finite dimension parameter in the space Θ. Thus the model marginal distribution of $\boldsymbol{x}$ is a mixture distribution with pdf $p(\boldsymbol{x}; \theta) = \sum_{k=1}^{g} \pi_k p(\boldsymbol{x}; \eta_k)$.

The parameter θ can be estimated by maximum likelihood (ml) from the training sample $\mathcal{D}$ by using the EM algorithm [7]. The straightforward formulas of EM in the semi-supervised context are not detailed here. They can be found for example in [15]. Denoting $\hat{\theta}_{\mathbf{x},\mathbf{z}}$ the maximum likelihood estimator (mle) of θ, a new observation $\boldsymbol{x}_{n+1}$ is assigned to one of the classes with the so-called maximum a posteriori (MAP) classification rule

$$r(\boldsymbol{x}_{n+1}; \hat{\theta}_{\mathbf{x},\mathbf{z}}) = \arg\max_{z \in \mathcal{Z}} p(z|\boldsymbol{x}_{n+1}; \hat{\theta}_{\mathbf{x},\mathbf{z}}). \tag{1}$$

2.2 Standard model selection criteria

Since many generative models can be used for a classification task, it is relevant to choose a model minimizing the expected misclassification error rate. Before addressing this question in the semi-supervised classification setting, some standard model selection criteria are now reviewed.

2.2.1 Cross-validated error rate

Cross-validation is a resampling technique which can be used to estimate the expected misclassification error rate $e(m) = E_{\mathbf{x},\mathbf{z}} E_{\boldsymbol{x},z}[\mathbf{1}_{\{r(\boldsymbol{x};\hat{\theta}^m_{\mathbf{x},\mathbf{z}})\neq z\}}]$ of the classifier derived from a model m with a training data set $(\mathbf{x},\mathbf{z})$. Here $\mathbf{1}$ denotes the indicator function, $(\boldsymbol{x},z)$ represent a new object to be classified and $\hat{\theta}^m_{\mathbf{x},\mathbf{z}}$ is the mle of the model parameter θ^m. The principle of V-fold cross-validation is the following:

1. Split at random $\mathcal{D}_u$ and $\mathcal{D}_\ell$ in V blocks of approximately equal sizes $\{\mathcal{D}_\ell^{\{1\}},\ldots,\mathcal{D}_\ell^{\{V\}}\}$ and $\{\mathcal{D}_u^{\{1\}},\ldots,\mathcal{D}_u^{\{V\}}\}$, respectively.

2. Denoting $\hat{\theta}^{\{-i\}}$ the mle derived from the training data $\mathcal{D}\backslash\{\mathcal{D}_\ell^{\{i\}},\mathcal{D}_u^{\{i\}}\}$ with the ith part of the data removed, compute

$$\hat{e}_i = \frac{1}{\operatorname{card}\left(\mathcal{D}_\ell^{\{i\}}\right)} \sum_{(\boldsymbol{x},z)\in\mathcal{D}_\ell^{\{i\}}} \mathbf{1}_{\{r(\boldsymbol{x};\hat{\theta}^{\{-i\}})\neq z\}} \quad (i=1,\ldots,V). \tag{2}$$

3. Define the estimation of the error rate of the classifier related to model m as $\hat{e}(m) = \frac{1}{V}\sum_{i=1}^{V}\hat{e}_i$.

It is important to note that a proper extension of cross-validation to the semi-supervised setting leads to remove the same proportion of labeled and unlabeled data from the training sample when computing $\hat{\theta}^{\{-i\}}$.

The model with the lowest cross-validated error rate $\hat{e}(m)$ is selected. This model selection criterion is expected to work well in most practical settings and is widely used. But as noticed in [12], V is a sensitive parameter to be chosen carefully to get a good error rate estimation. In the experiments presented in Sections 4 and 5, cross-validation with $V=3$ (CV_3) and $V=10$ (CV_{10}) has been considered. Note also that cross-validation can be computationally demanding for semi-supervised classification since it needs to run V EM algorithms for estimating $\hat{\theta}^{\{-1\}},\ldots,\hat{\theta}^{\{-V\}}$.

2.2.2 AIC, a deviance criterion

AIC is a general model selection criterion which is not estimating directly the expected misclassification error rate but is cheaper than cross-validation. It consists of an asymptotic approximation of the expected generative deviance of the model m

$$\Delta(m) = 2E_{\mathbf{x},\mathbf{z},\mathbf{x}',\mathbf{z}'}[\log p(\mathbf{x}',\mathbf{z}') - \log p(\mathbf{x}',\mathbf{z}';\hat{\theta}^m_{\mathbf{x},\mathbf{z}})], \tag{3}$$

where $\mathbf{x},\mathbf{z}$ and $\mathbf{x}',\mathbf{z}'$ are two independent samples. It leads to the following criterion:

$$\mathrm{AIC}(m) = 2\log p(\mathbf{x},\mathbf{z};\hat{\theta}^m_{\mathbf{x},\mathbf{z}}) - 2\nu_m, \tag{4}$$

where ν_m is the dimension of the parameter space for model m. The model with the largest AIC value is selected.

For nested true models, AIC can select with non zero probability overfitted models, so that AIC is not consistent. In a regression framework AIC has been proved to be optimal for the minimax risk [10]. In a classification context, AIC could be expected to choose a model with a low error rate but depending of a too complex parameter θ.

2.2.3 BIC, a Bayesian information criterion

BIC arises in a Bayesian approach of model selection. It consists of an asymptotic approximation of the log-integrated likelihood of model m:

$$\log p(\mathbf{x}, \mathbf{z}|m) = \log \int_{\Theta^m} p(\mathbf{x}, \mathbf{z}; \theta^m) p(\theta^m) d\theta^m, \tag{5}$$

$p(\theta^m)$ being a prior distribution on the parameter θ^m in the parameter space Θ^m. It leads to the criterion

$$\mathrm{BIC}(m) = \log p(\mathbf{x}, \mathbf{z}; \hat{\theta}^m_{\mathbf{x},\mathbf{z}}) - \frac{\nu_m}{2} \log n, \tag{6}$$

which does not depend on the prior distribution $p(\theta^m)$. The model with the largest BIC value is selected. BIC is a consistent criterion: it selects the less complex true model with probability 1 as n grows to infinity under regularity conditions. Note that BIC is not focussing on the classification task, but as for AIC, it can be expected to select a good classifier as soon as a good approximation of the joint distribution of the data is provided by at least one model.

2.2.4 BEC, a Bayesian entropic criterion

Since a good classifier relies on a good approximation of the conditional distribution $p(\mathbf{z}|\mathbf{x}; \hat{\theta}^m_{\mathbf{x},\mathbf{z}})$ (see Equation (1)), it is sensible to choose a generative classification model providing the largest conditional integrated likelihood $p(\mathbf{z}|\mathbf{x}, m)$. In this Bayesian perspective, the BEC criterion to be maximized is a BIC-like approximation of $\log p(\mathbf{z}|\mathbf{x}, m)$:

$$\mathrm{BEC}(m) = \log p(\mathbf{x}, \mathbf{z}; \hat{\theta}^m_{\mathbf{x},\mathbf{z}}) - \log p(\mathbf{x}; \hat{\theta}^m_{\mathbf{x}}), \tag{7}$$

where $\hat{\theta}^m_{\mathbf{x}}$ is the mle of θ^m derived from $\mathbf{x}$. The computational cost of BEC is larger than AIC or BIC since both $\hat{\theta}^m_{\mathbf{x},\mathbf{z}}$ and $\hat{\theta}^m_{\mathbf{x}}$ have to be estimated through an EM algorithm, but it remains by far cheaper than cross-validation.

From a theoretical point of view, if the sampling distribution belongs to a single model of the model collection, this model will be asymptotically selected by BEC [4]. However, when there are several nested true models, BEC can select arbitrarily complex models among them. From a practical point of view, BEC has been proved to behave better than AIC and BIC for many classification problems but often selects more complex generative classifiers than the cross-validated error rate criterion [4].

3 A predictive deviance criterion

We propose a new criterion for selecting a classifier in the semi-supervised setting. This criterion aims at selecting a model producing good performances in classification with a computational cost smaller than cross-validation criteria. In a frequentist perspective, a quantity of interest to select a generative classifier with good prediction performances is the predictive deviance of the classification model which is related to the conditional likelihood of the model knowing the predictors. Thus, we are aiming to find the model minimizing the

expected Kullback-Leibler (KL) divergence between the estimated conditional distribution of $z|\boldsymbol{x}$ and the true conditional distribution:

$$2\mathbb{E}_{\mathbf{x},\mathbf{z},\mathbf{x}',\mathbf{z}'}[\log p(\mathbf{z}'|\mathbf{x}') - \log p(\mathbf{z}'|\mathbf{x}';\hat{\theta}^m_{\mathbf{x},\mathbf{z}})], \tag{8}$$

with $(\mathbf{x},\mathbf{z})$ and $(\mathbf{x}',\mathbf{z}')$ two independent samples. Since the first term does not depend on the model, it leads to find the model maximizing:

$$E_{cond}(m) = 2\mathbb{E}_{\mathbf{x},\mathbf{z},\mathbf{x}',\mathbf{z}'} \log p(\mathbf{z}'|\mathbf{x}';\hat{\theta}^m_{\mathbf{x},\mathbf{z}}). \tag{9}$$

Proposition 1 below provides an estimation of $E_{cond}(m)$ under the hypothesis (H1) that there is a true model m:

(H1): it exists $\theta_0^m \in \Theta^m$ such that for all $(\boldsymbol{x},z) \in \mathcal{X} \times \mathcal{Z}$, $p(\boldsymbol{x},z) = p(\boldsymbol{x},z;\theta_0^m)$.

The proof of Proposition 1 is given in Appendix A.

Proposition 1 *If the model m verifies (H1) and under standard regularity conditions we have*

$$E_{cond}(m) = 2[\log p(\mathbf{x},\mathbf{z};\hat{\theta}^m_{\mathbf{x},\mathbf{z}}) - \log p(\mathbf{x};\hat{\theta}^m_{\mathbf{x}})] - [\nu_m - tr(JJ_\beta^{-1})] + O_p(\sqrt{n}), \tag{10}$$

J and J_β being respectively the Fisher information matrices for unlabeled and partially-labeled data: $J = -\mathbb{E}_{\boldsymbol{x}}[\nabla^2 \log p(\boldsymbol{x};\theta_0^m)]$ and $J_\beta = \beta J_c + (1-\beta)J$, where $J_c = -\mathbb{E}_{\boldsymbol{x},z}[\nabla^2 \log p(\boldsymbol{x},z;\theta_0^m)]$.

Equation (10) exhibits a specific penalty $[\nu_m - \mathrm{tr}(JJ_\beta^{-1})]$ which deserves comments. This penalty depends on the class overlapping and can be related to the number of predictive parameters present in the generative model. First, It can be noticed that when classes are well-separated, $J \approx J_c$ and consequently $J \approx J_\beta$ so that $\nu_m - \mathrm{tr}(JJ_\beta^{-1}) \approx 0$. On the contrary, the more the classes are overlapping, the more $(\nu_m - \mathrm{tr}(JJ_\beta^{-1}))$ is large. This claim can be made precise in particular Gaussian situations (see [21]). It is illustrated in the following example:

Suppose that data are generated according to $X|Z=1 \sim \mathcal{N}(0,1)$, $X|Z=2 \sim \mathcal{N}(\Delta,1)$ and $\pi_1 = \pi_2 = 0.5$. Figure 1 reports the value of the penalty according to Δ for an heteroscedastic Gaussian model in the supervised setting ($\beta = 1$). The penalty is maximum when the classes are not separated. It is important to notice that when $\Delta = 0$, the penalty is equal to the number of parameters involved into the quadratic logistic regression.

The penalty $(\nu_m - \mathrm{tr}(JJ_\beta^{-1}))$ is difficult to derive since it requires the computation of the information matrices in a mixture framework. Proposition 2, proved in Appendix B, provides a simple mean to approximate it.

Proposition 2 *If the model m verifies (H1) and under standard regularity conditions then*

$$[\nu_m - tr(JJ_\beta^{-1})] = 2(\log p(\mathbf{x};\hat{\theta}^m_{\mathbf{x}}) - \log p(\mathbf{x};\hat{\theta}^m_{\mathbf{x},\mathbf{z}})) + O_p(\sqrt{n}). \tag{11}$$

It leads to the following expression for $E_{cond}(m)$:

$$E_{cond}(m) = 2\log p(\mathbf{z}|\mathbf{x};\hat{\theta}^m_{\mathbf{x},\mathbf{z}}) - 4[\log p(\mathbf{x};\hat{\theta}^m_{\mathbf{x}}) - \log p(\mathbf{x};\hat{\theta}^m_{\mathbf{x},\mathbf{z}})] + O_p(\sqrt{n}). \tag{12}$$

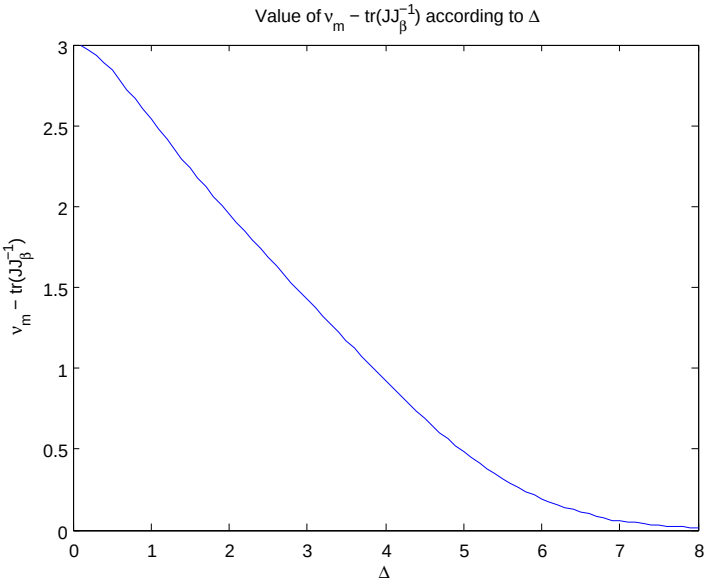


Figure 1: Value of the penalty according to the class separation.

and to the criterion

$$\mathrm{AIC}_{cond}(m) = 2\log p(\mathbf{z}|\mathbf{x};\hat{\theta}^m_{\mathbf{x},\mathbf{z}}) - 4\log\frac{p(\mathbf{x};\hat{\theta}^m_{\mathbf{x}})}{p(\mathbf{x};\hat{\theta}^m_{\mathbf{x},\mathbf{z}})}. \tag{13}$$

The approximation error centered at zero that AIC_{cond} involves is relatively high (in $O_p(\sqrt{n})$) as for AIC.

Note that AIC_{cond} can be viewed as an overpenalized BEC criterion since it can be written

$$\mathrm{AIC}_{cond}(m) = 2\mathrm{BEC}(m) - 2\log\frac{p(\mathbf{x};\hat{\theta}^m_{\mathbf{x}})}{p(\mathbf{x};\hat{\theta}^m_{\mathbf{x},\mathbf{z}})}. \tag{14}$$

The additional penalty is expected to avoid that a plateau appears when considering true nested models as shown in the following proposition (proved in Appendix C).

Proposition 3 *Assume the data distribution belongs to two nested models m and m' with $m \subset m'$. If the number of data is large enough:*

$$\mathbb{E}[AIC_{cond}(m)] - \mathbb{E}[AIC_{cond}(m')] > 0. \tag{15}$$

In other words, AIC_{cond} tends to prefer the less complex model between two nested right models. Moreover, as BEC, AIC_{cond} is selecting the right model when it is unique, as specified in the next proposition.

Proposition 4 *If the sampling distribution belongs to one and only one model m^* in a finite collection $\{m_1, ..., m_M\}$, and under standard regularity conditions on the parametric family, then AIC_{cond} selects m^*.*

The proof of this proposition is analogous to the proof of Proposition 1 in [4] for BEC criterion and is omitted.

The criterion AIC_{cond} can be thought of as a promising model selection criterion for classification in the semi-supervised context. In this context, it does not require additional computations compared to criteria as AIC or BIC since the EM algorithm is run to improve the classifier by taking into account the numerous unlabeled data. The focus of AIC_{cond} is prediction and, moreover, from Proposition 3, it is expected to be more parsimonious than BEC. Finally, AIC_{cond} can be expected to choose a reliable and parsimonious classifier. Its practical behavior is studied in Sections 4 and 5 from numerical experiments on simulated and real data sets where it is compared to the criteria presented in Section 2.2.

4 Experiments on simulated data

4.1 Variable selection

The first numerical experiment concerns a simple variable selection problem in the Gaussian setting. This experiment aims to contrast the behavior of AIC_{cond} and BEC for a problem for which the true model is not providing the lowest error rate.

Data have been simulated according to a design where the variables bring less and less information to finally deteriorate the classification error rate. The experimental setting is as follows: $g = 2$, $\pi_1 = \pi_2 = 0.5$ and the class-conditional distributions are Gaussian distributions, $\boldsymbol{x}|z = 1 \sim \mathcal{N}(0_{50\times 1}, I_{50})$ and $\boldsymbol{x}|z = 2 \sim \mathcal{N}(\mu, I_{50})$ with $\mu_i = \frac{1}{i}\ \forall i \in \{1, \ldots, 50\}$, and I_d stands for the identity matrix of dimension d. Thus, variables provide less and less discriminant information. The order in which variables are selected from 1 to 50 is assumed to be known. The true model leads to select all the variables, but the less informative variables increase dramatically the classifier variance. A test sample of size 50 000 has been simulated. Four combinations of n_ℓ labeled data and n_u unlabeled data have been considered: S_1 with $n_\ell = 100$ and $n_u = 0$; S_2 with $n_\ell = 1\,000$ and $n_u = 0$; SS_1 with $n_\ell = 100$ and $n_u = 1\,000$; SS_2 with $n_\ell = 1\,000$ and $n_u = 10\,000$. Each combination has been replicated 100 times.

The optimal, actual and apparent error rates according to the number of selected variables are shown for SS_1 in Figure 2. The optimal and apparent error rates decrease as the number of selected variables increases, while the actual error rate on the test sample decreases and then increases.

	BEC	AIC_{cond}	CV_3	CV_{10}	NbVar*
S_1	10.5	**3.1**	7.8	7.8	3
S_2	21.7	**11.3**	12.2	14.0	11
SS_1	17.5	**9.2**	10.7	10.0	6
SS_2	33.8	**22.0**	21.1	21.4	23

Table 1: Variable selection for simulated data: Average number of selected variables for each criterion (best criterion in bold).

For this experiment, the performances of CV_3, CV_{10}, BEC and AIC_{cond} criteria have been compared. The results are summarized in Tables 1 and 2

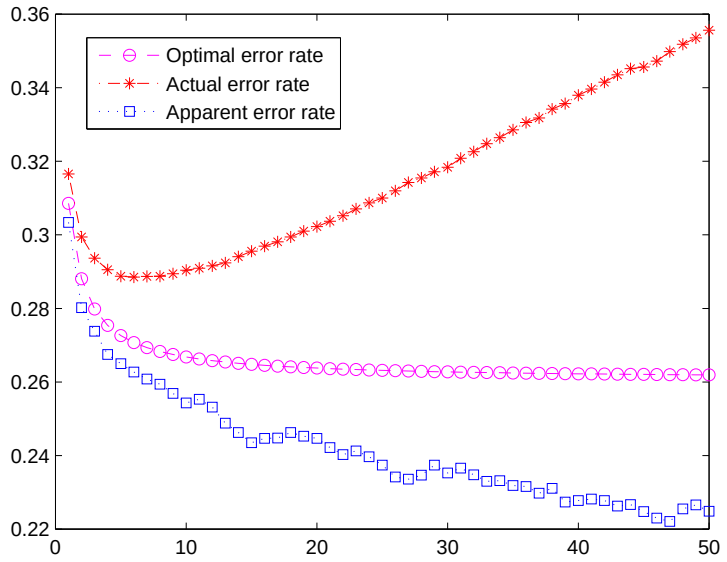


Figure 2: Variable selection for simulated data: Error rates according to the number of selected variables with design SS_1.

	BEC	AIC_{cond}	CV_3	CV_{10}	Err^*
S_1	31.53	**30.40**	31.08	31.08	29.68
S_2	27.90	**27.68**	27.77	27.78	27.55
SS_1	30.42	29.75	**29.70**	29.82	28.55
SS_2	27.18	**27.17**	**27.17**	27.21	27.03

Table 2: Variable selection for simulated data: Error rate (%) for the different criteria (best criterion in bold).

where NbVar^* denotes the optimal number of variables derived from the actual error rate function and Err^* the corresponding error rate. Those tables show that AIC_{cond} performs the best since it selects in average the most accurate number of variables (Table 1) and produces a low classification error rate in both the supervised and semi-supervised settings (Table 2). Cross-validation also produces good results in both settings and BEC behaves poorly because it selects too many variables. This experiment shows that for nested reliable models, AIC_{cond} leads to choose a parsimonious model with good prediction performances contrary to BEC.

4.2 Choosing a MDA model

Mixture Discriminant Analysis (MDA) is a versatile generative classification model [11] which consists of modeling each class-conditional distribution with a mixture of Gaussian distributions. The interest of MDA in the semi-supervised setting, where the availability of many unlabeled data help to estimate properly the marginal distribution on $\boldsymbol{x}$, has already been pointed out in [15]. In MDA, sensitive parameters to be fixed are the number of mixture components per class.

Numerical experiments have been performed to compare the behavior of AIC_{cond} with CV_3, CV_{10}, AIC, BIC and BEC to select those discrete parameters. To avoid combinatorial issues, each class-conditional distribution is assumed to have the same number of mixture components.

A two-class problem with three mixture components per class is considered. Each component C_i is following a $\mathcal{N}(\mu_i, 0.15I_6)$ distribution with $\mu_i = (\mu_{i1}, \mu_{i2}, 0, 0, 0, 0)'$ and the mixing proportions are equal. Class 1 distribution is a mixture of the three components C_1, C_2 and C_3 with $(\mu_{11}, \mu_{12}) = (0, 0)$, $(\mu_{21}, \mu_{22}) = (1, 1)$ and $(\mu_{31}, \mu_{32}) = (2, 0)$. Class 2 distribution is a mixture of the three components C_4, C_5 and C_6 with $(\mu_{41}, \mu_{42}) = (1, 0)$, $(\mu_{51}, \mu_{52}) = (2, -1)$ and $(\mu_{61}, \mu_{62}) = (3, 0)$. Moreover, the prior probabilities of Classes 1 and 2 are equal. The simulated model is depicted in Figure 3. We simulated 100 independent samples with $n_\ell = 100$ and $n_u = 1\ 000$ and a test sample of size 50 000. The considered models were heteroscedastic Gaussian mixture models with diagonal variance matrices with one to five components per class. The number of times each model has been chosen by BIC, AIC, BEC, AIC_{cond}, CV_3 and CV_{10} is reported in Table 3 which also provides the mean error rate (%) for each model.

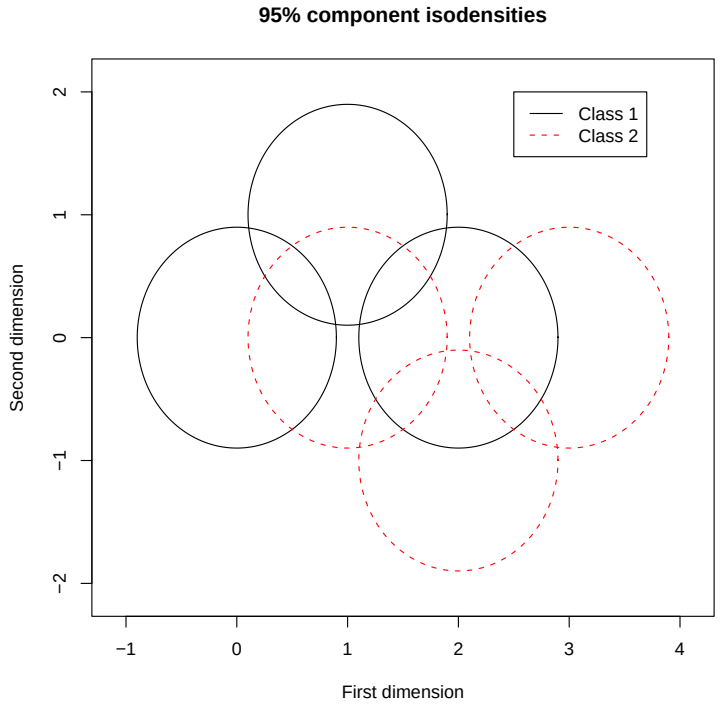


Figure 3: MDA model selection for simulated data: 95% component isodensities of each component.

In this table, CC denotes the chosen number of mixture component per class. The right choice is $CC = 3$. It can be seen that BIC performs the best, often choosing the right number of components. BEC tends to select too complex

CC	BIC	AIC	BEC	AIC_{cond}	CV_3	CV_{10}	$\bar{Err}$
1	0	0	0	0	0	0	26.22
2	0	0	0	0	3	2	22.42
3*	97	64	32	53	45	42	**15.61**
4	3	15	31	27	25	31	16.07
5	0	21	37	20	27	25	16.62
$\bar{Err}$	**15.80**	15.97	16.08	15.93	16.07	16.07	15.60

Table 3: MDA model selection for simulated data: Number of components per class (CC) selected by each criterion (criterion producing the lowest error rate is in bold).

models, as AIC_{cond} and AIC to a smaller extent. CV_3 and CV_{10} are quite spread over all possible models. In this example the best criterion is, by far, BIC followed by AIC_{cond}. This outstanding behavior of BIC is not a surprise since in practical situations this criterion is known to select the right number of components of mixture when the data actually arose from a mixture as soon as the sample size is reasonable (see for instance [9]). However, we will see hereafter that, contrary to AIC_{cond}, the behavior of BIC on real data sets could be quite disappointing.

4.3 Choosing the class variance matrices

Eigenvalue decomposition of the class conditional variance matrices Σ_k, $k = 1, \ldots, g$ allows to define many parsimonious Gaussian classification models including LDA and QDA [5]. Those models allow to interpret class variance matrices in terms of volume, orientation and shape of the classes and have shown a good behavior on many real datasets. Usually, a model is chosen by minimizing the cross-validated error rate. Here we study the behavior of the above mentioned criteria to select a model among the homoscedastic model with a common spherical variance matrix ($\Sigma_k = \lambda I$), the homoscedastic model with a common diagonal variance matrix ($\Sigma_k = \lambda B$), the standard homoscedatic model ($\Sigma_k = \lambda C$), the heteroscedastic model with spherical variance matrices ($\Sigma_k = \lambda_k I$), the heteroscedastic model with diagonal variance matrices ($\Sigma_k = \lambda_k B_k$) and the standard heteroscedastic model ($\Sigma_k = \lambda_k C_k$).

We simulated 100 independent training samples with $n_u = 2\,000$, $n_\ell = 200$ and a test sample of size 50 000 from a two class model. Classes are generated according to $\mathcal{N}(\mu_k, \Sigma_k)$ where Σ_k is a diagonal matrix with $\Sigma_1 = 2 \times \text{diag}(2, 1.5, 1, 1, 1, 1)$ and $\Sigma_2 = \text{diag}(2, 1.5, 1, 1, 1, 1)$, $\mu_1 = (2, 0, 0, 0, 0, 0)'$, $\mu_2 = (0, 0, 0, 0, 0, 0)'$ and $\pi_1 = \pi_2 = 0.5$. Table 4 displays the number of times each criterion selects a model and the average error rate (%) produced by each model. The last line provides the error rates produced by each criterion and the minimal error rate. BEC and AIC_{cond} have the same behavior, however AIC_{cond} selects more parsimonious models. BIC and AIC almost always select the less complex true model and consequently produce very good results. CV_{10} and CV_3 do not work so well on this instance. This is probably due to the excess of variability of the estimated classification error rate. Those experiments show that under true model assumptions information-based model selection criteria perform better than cross-validation. On this extent, the best criteria are BIC and AIC. However, BEC and AIC_{cond} perform well in comparison to

	BIC	AIC	BEC	AIC_{cond}	CV_3	CV_{10}	$\bar{Err}$
λI	0	0	0	0	1	0	27.49
$\lambda_k I$	0	0	1	1	98	41	22.97
λ_B	0	0	0	0	0	0	27.80
$\lambda_k B*$	100	98	49	62	1	34	**20.60**
λC 0	0	0	0	0	0	0	28.34
$\lambda_k C$	0	2	50	37	0	25	20.66
Err^*	**20.60**	**20.60**	20.67	20.66	23.00	21.61	20.58

Table 4: Class variance matrices selection for simulated data: Number of times a model is selected by each criterion (criterion producing the lowest error rate is in bold).

cross-validation. In particular they produce lower error rates. We will see on the experiments on real data below that BEC and AIC_{cond} still perform well contrary to BIC and AIC.

5 Experiments on real data

In this section, the criteria behavior is compared on some benchmark datasets from the UCI database repository[1] and Pattern Recognition datasets[2] and on a dataset coming from [3] on seabirds sex prediction.

5.1 Experiments on benchmark datasets

The performances of criteria for selecting a model among the six models described in Section 4.3 are studied on some of the UCI and Pattern Recognition datasets. Features of the datasets are summarized in Table 5. If a test set is provided, its predictors are used to learn the parameters of the classification models in the semi-supervised setting and its labels are used to compute the error rate. Otherwise 100 random splits of n_u unlabeled data and n_ℓ labeled data are generated. Table 6 shows that AIC_{cond}, BEC and cross-validation have a similar behavior and outperform BIC and AIC as in the Parkinson and Pima datasets.

Dataset	n	d	g	Test set	n_u	n_ℓ
Crab	200	5	4	no	150	50
Iris	150	4	3	no	100	50
Parkinson	195	22	2	no	95	100
Pima	532	7	2	yes	332	200
Wine	178	13	3	no	89	89

Table 5: Variable parameterization selection for benchmark datasets: Experimental setting.

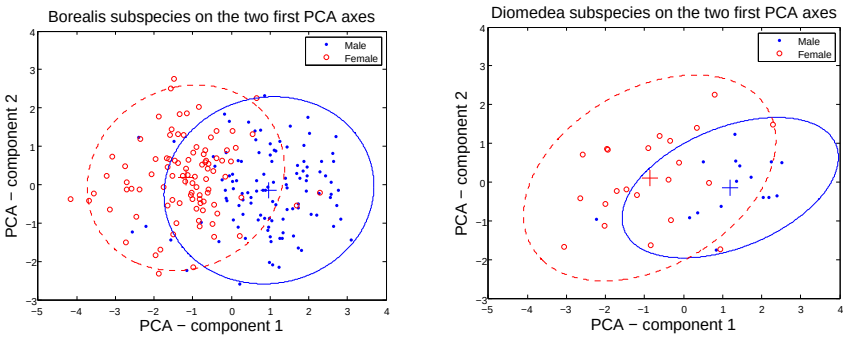
[1] http://archive.ics.uci.edu/ml/
[2] http://www.stats.ox.ac.uk/pib/PRNN/

	BIC	AIC	BEC	AIC_{cond}	CV_3	CV_{10}
Crab	**6.63**	6.75	6.80	6.77	7.81	7.78
Iris	2.98	2.98	**2.91**	**2.91**	3.25	3.21
Parkinson	26.45	30.68	15.43	**15.16**	18.20	16.38
Pima	25.00	25.00	**19.58**	**19.58**	22.53	**19.58**
Wine	3.24	**1.17**	1.45	1.47	1.73	1.70

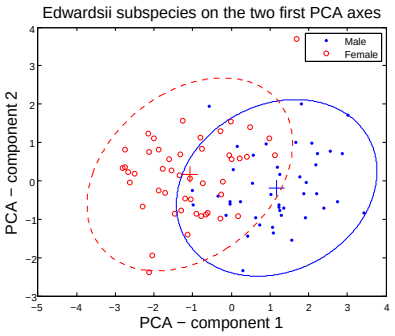
Table 6: Variable parameterization selection for benchmark datasets: Error rate of each criterion on UCI datasets (criterion producing the lowest error rate is in bold).

5.2 Seabirds dataset

We consider a dataset coming from [3] on seabirds from the family Procellaridae (petrels). The problem is to predict the gender of the 336 birds from five continuous predictors (culmen (bill length), tarsus, wing and tail lengths, and culmen depth). Three sub-species (*borealis*, *diomedea* and *edwardsii*) are considered. After centering and reducing the data subspecies-wise, each subspecies can be considered as coming from the same distribution. This standard normalization is in accordance with biological assumptions and, as shown in Figure 4, it leads to get similar distributions for each sub-species.



"unlabeled" learning data labeled learning data



test data

Figure 4: Normalized petrels data and 95% isodensities for each class.

We consider the smallest *diomedea* subspecies (38 birds) as labeled data, the *borealis* (206 birds) subspecies is considered as unlabeled data and the *edwardsii* subspecies (92 birds) is considered as test data. This provides a natural partitioning between labeled, unlabeled and test data. We consider the model selection issue in the supervised setting (only *diomedea* in the learning sample) and in the semi-supervised setting (both *diomedea* and *borealis* in the learning sample), the goal being to select one of the 28 model proposed in [5]. The results of this experiment are shown in Tables 7 and 8 where the classification error rate on the test sample *edwardsii* and the model selected by each criterion are displayed.

Model	Error (%)	Criteria
$\pi\lambda I$	15.22	BEC, CV_3, CV_{10}
$\pi\lambda_k I$	13.04	AIC_{cond}, CV_{10}
$\pi_k\lambda I$	14.13	CV_3
$\pi\lambda B$	13.04	CV_3
$\pi\lambda_k B$	**8.70**	$\mathbf{CV}_3$
$\pi\lambda B_k$	10.87	CV_{10}
$\pi_k\lambda B$	10.87	CV_3
$\pi_k\lambda B_k$	14.13	CV_{10}
$\pi\lambda C$	10.87	CV_{10}
$\pi\lambda D A_k D$	19.57	BIC, AIC
$\pi_k\lambda C$	11.96	CV_{10}

Table 7: Error rate for the test sample *edwardsii* produced by each model selected by at least one criterion, and the error rate produced by the best model, in the supervised setting.

Model	Error (%)	Criteria
$\pi\lambda I$	16.30	CV_3, CV_{10}
$\pi\lambda_k I$	14.13	CV_3, CV_{10}
$\pi\lambda B$	11.96	CV_3, CV_{10}
$\pi\lambda_k B$	11.96	CV_3, CV_{10}
$\pi\lambda B_k$	13.04	CV_3, CV_{10}
$\pi\lambda_k B_k$	11.96	CV_3, CV_{10}
$\pi_k\lambda B$	13.04	CV_3, CV_{10}
$\pi_k\lambda_k B$	11.96	CV_3, CV_{10}
$\pi_k\lambda B_k$	11.96	CV_3, CV_{10}
$\pi_k\lambda_k B_k$	9.78	CV_3, CV_{10}
$\pi\lambda C$	11.96	CV_3, CV_{10}
$\pi\lambda_k C$	11.96	CV_{10}
$\pi\lambda C_k$	8.70	$\mathbf{CV}_3$
$\pi_k\lambda_k D A_k D$	8.70	$\mathbf{AIC}_{cond}$
$\pi_k\lambda_k D_k A D_k$	9.78	BEC
$\pi_k\lambda C_k$	9.78	AIC,BIC
$\pi\lambda_k C_k$	**7.61**	

Table 8: Error rate for the test sample *edwardsii* produced by each model selected by at least one criterion, and the error rate produced by the best model, in the semi-supervised setting.

First, it is noteworthy that using the unlabeled sample *borealis* leads to a great improvement in the classification error. Second, it is striking that cross-validation leads to many ties and, for this very reason, appears to be no very helpful to choose a proper model. For instance, selecting the simplest model providing the smallest cross-validated error rate is disappointing. It produces a test error rate of 15.2 % (resp. 16.30 %) in the supervised (resp. semi-supervised) setting. Moreover, no surprisingly the choice of V in the V-fold cross-validation appears to be sensitive. For the present example choosing $V = 3$ has to be preferred to $V = 10$, but there is no a priori reason to favor $V = 3$ against $V = 10$ or other values,... Finally, this example highlights the good behavior of AIC_{cond} which outperforms the other model selection criteria.

6 Discussion

We have proposed a predictive deviance criterion, called AIC_{cond}, devoted to selected a generative classification model in the semi-supervised setting. This criterion has been conceived to select a parsimonious generative classifier leading to a low misclassification error rate. To this end, AIC_{cond} is aiming to minimize the expected KL distance between the model-wise estimated conditional distribution of the labels knowing the predictors and the true conditional distribution. Derived from asymptotic approximations of the conditional deviance of the model, AIC_{cond} is comparing in some sense the ml parameter estimate derived from the labeled data set and the ml parameter estimate derived from the unlabeled data set. This approach appears quite natural in a semi-supervised context where it is possible to use the EM algorithm to take profit of the unlabeled data to improve the classifier. From the practical point of view, in the present context, the well-documented drawbacks of the EM algorithm which are high dependence on initial position, slow convergence, and the existence of spurious local maximizers are not expected to jeopardize the computation of criterion AIC_{cond}. Actually, the EM algorithm can be initialized in a quite natural way with the parameter ml estimate derived from the labeled data. Thus, convergence towards insensible fixed point of the EM algorithm are not expected to occur.

The criterion AIC_{cond} provides an efficient alternative to the cross-validated error rate when the collection of models under consideration is large. This criterion which is an AIC-like approximation of the expected predictive deviance provided by a generative model, leads in many cases to select a model with a lower error rate than the AIC and BIC criteria.

Formally, AIC_{cond} appears as a reminiscent of BEC criterion with an additional penalty which ensures consistency over nested models. Thus, AIC_{cond} is answering a possible drawback of BEC which could lead to select too complex models when nested models are in competition. Finally, as illustrated with numerical experiments on simulated and real datasets, AIC_{cond} has to be preferred to BEC: In many situations they perform the same, but when they give different answers, AIC_{cond} outperforms BEC.

A Proof of Proposition 1

For the sake of simplicity the model m does not index the formulas. We need first the following lemma.

Lemma 1 *Under (H1) and noting θ_0 the true value of θ, we have*

$$\sqrt{n}(\hat{\theta}_{\mathbf{x},\mathbf{z}} - \theta_0) \xrightarrow{\mathcal{D}} \mathcal{N}(0, J_\beta^{-1}). \tag{16}$$

Proof of Lemma 1

(H1) ensures that $\hat{\theta}_{\mathbf{x},\mathbf{z}} \xrightarrow{P} \theta_0$. A first order Taylor expansion about θ_0 gives

$$\nabla \log p(\mathbf{x}, \mathbf{z}; \hat{\theta}_{\mathbf{x},\mathbf{z}}) = 0 = \nabla \log p(\mathbf{x}, \mathbf{z}; \theta_0) + \nabla^2 \log p(\mathbf{x}, \mathbf{z}; \bar{\theta})(\hat{\theta}_{\mathbf{x},\mathbf{z}} - \theta_0) \tag{17}$$

with $\bar{\theta}$ on the line joining θ_0 and $\hat{\theta}_{\mathbf{x},\mathbf{z}}$. Thus

$$\sqrt{n}(\hat{\theta}_{\mathbf{x},\mathbf{z}} - \theta_0) = \left[-\frac{1}{n}\nabla^2 \log p(\mathbf{x},\mathbf{z};\bar{\theta})\right]^{-1} \frac{1}{\sqrt{n}}\nabla \log p(\mathbf{x},\mathbf{z};\theta_0). \tag{18}$$

Owing to the law of large numbers and standard regularity conditions

$$\left[-\frac{1}{n}\nabla^2 \log p(\mathbf{x},\mathbf{z};\bar{\theta})\right]^{-1} \xrightarrow{P} J_\beta^{-1}, \tag{19}$$

where $J_\beta = \beta J_c + (1-\beta)J$, with $J_c = -\mathbb{E}_{\boldsymbol{x},z}[\nabla^2 \log p(\boldsymbol{x},z;\theta_0)]$ and $J = -\mathbb{E}_{\boldsymbol{x}}[\nabla^2 \log p(\boldsymbol{x};\theta_0)]$. Standard central limit theorem cannot be directly applied on $\frac{1}{\sqrt{n}}\nabla \log p(\mathbf{x},\mathbf{z};\theta_0)$ since $\nabla \log p(\mathbf{x},\mathbf{z};\theta_0)$ is not a sum of iid variables, however it can be applied on labeled data and on unlabeled data to lead to

$$\frac{1}{\sqrt{n}}\nabla \log p(\mathbf{x},\mathbf{z};\theta_0) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \beta K_c + (1-\beta)K), \tag{20}$$

with $K_c = \mathrm{Var}[\nabla \log p(\boldsymbol{x},z;\theta_0)]$ and $K = \mathrm{Var}[\nabla \log p(\boldsymbol{x};\theta_0)]$. Now the true model assumption (H1) ensures that $K_c = J_c$ and $K = J$ and then

$$\sqrt{n}(\hat{\theta}_{\mathbf{x},\mathbf{z}} - \theta_0) \xrightarrow{\mathcal{D}} \mathcal{N}(0, J_\beta^{-1}). \tag{21}$$

Proof of Proposition 1

We can now prove Proposition 1. We start by establishing the following equations:

$$\begin{aligned}
2\mathbb{E}_{\mathbf{x},\mathbf{z},\mathbf{x}',\mathbf{z}'}[\log p(\mathbf{x}',\mathbf{z}';\hat{\theta}_{\mathbf{x},\mathbf{z}})] &= 2\mathbb{E}_{\mathbf{x},\mathbf{z}}[\log p(\mathbf{x},\mathbf{z};\theta_0)] - \nu + o(1) &(22)\\
2\mathbb{E}_{\mathbf{x},\mathbf{z}}[\log p(\mathbf{x},\mathbf{z};\theta_0)] &= 2\mathbb{E}_{\mathbf{x},\mathbf{z}}[\log p(\mathbf{x},\mathbf{z};\hat{\theta}_{\mathbf{x},\mathbf{z}})] - \nu + o(1) &(23)\\
2\mathbb{E}_{\mathbf{x},\mathbf{z},\mathbf{x}'}[\log p(\mathbf{x}';\hat{\theta}_{\mathbf{x},\mathbf{z}})] &= 2\mathbb{E}_{\mathbf{x}}[\log p(\mathbf{x};\theta_0)] - \mathrm{tr}(JJ_\beta^{-1}) + o(1) &(24)\\
2\mathbb{E}_{\mathbf{x}}[\log p(\mathbf{x};\theta_0)] &= 2\mathbb{E}_{\mathbf{x}}[\log p(\mathbf{x};\hat{\theta}_{\mathbf{x}})] - \nu + o(1). &(25)
\end{aligned}$$

Equations (23) and (25) state the same result.

We first prove Equation (22). By a Taylor expansion about θ_0, we obtain:

$$\begin{aligned}
2\log p(\mathbf{x}',\mathbf{z}';\hat{\theta}_{\mathbf{x},\mathbf{z}}) &= 2\log p(\mathbf{x}',\mathbf{z}';\theta_0) + 2(\hat{\theta}_{\mathbf{x},\mathbf{z}} - \theta_0)'\nabla_\theta \log p(\mathbf{x}',\mathbf{z}';\theta_0) &(26)\\
&\quad +(\hat{\theta}_{\mathbf{x},\mathbf{z}} - \theta_0)'\nabla^2_\theta \log p(\mathbf{x}',\mathbf{z}';\bar{\theta})(\hat{\theta}_{\mathbf{x},\mathbf{z}} - \theta_0) &(27)\\
&= 2\log p(\mathbf{x}',\mathbf{z}';\theta_0) + 2(\hat{\theta}_{\mathbf{x},\mathbf{z}} - \theta_0)'\nabla_\theta \log p(\mathbf{x}',\mathbf{z}';\theta_0) &(28)\\
&\quad +\mathrm{tr}[\nabla^2_\theta \log p(\mathbf{x}',\mathbf{z}';\bar{\theta})(\hat{\theta}_{\mathbf{x},\mathbf{z}} - \theta_0)(\hat{\theta}_{\mathbf{x},\mathbf{z}} - \theta_0)'] &(29)
\end{aligned}$$

with $\bar{\theta}$ on the line joining θ_0 and $\hat{\theta}_{\mathbf{x},\mathbf{z}}$. Taking the expectation we get

$$\begin{aligned}
2\mathbb{E}_{\mathbf{x},\mathbf{z},\mathbf{x}',\mathbf{z}'}[\log p(\mathbf{x}',\mathbf{z}';\hat{\theta}_{\mathbf{x},\mathbf{z}})] &= \mathbb{E}_{\mathbf{x},\mathbf{z},\mathbf{x}',\mathbf{z}'}[\mathrm{tr}[\nabla^2_\theta \log p(\mathbf{x}',\mathbf{z}';\bar{\theta})(\hat{\theta}_{\mathbf{x},\mathbf{z}} - \theta_0)(\hat{\theta}_{\mathbf{x},\mathbf{z}} - \theta_0)']]\\
&\quad +2\mathbb{E}_{\mathbf{x},\mathbf{z},\mathbf{x}',\mathbf{z}'}[\log p(\mathbf{x}',\mathbf{z}';\theta_0)] + 2\mathbb{E}_{\mathbf{x},\mathbf{z},\mathbf{x}',\mathbf{z}'}[(\hat{\theta}_{\mathbf{x},\mathbf{z}} - \theta_0)'\nabla_\theta \log p(\mathbf{x}',\mathbf{z}';\theta_0)].
\end{aligned}$$

Since $(\mathbf{x}',\mathbf{z}')$ is an independent replicate of $(\mathbf{x},\mathbf{z})$, $\mathbb{E}_{\mathbf{x},\mathbf{z},\mathbf{x}',\mathbf{z}'}[\log p(\mathbf{x}',\mathbf{z}';\theta_0)] = \mathbb{E}_{\mathbf{x},\mathbf{z}}[\log p(\mathbf{x},\mathbf{z};\theta_0)]$ and $\mathbb{E}_{\mathbf{x},\mathbf{z},\mathbf{x}',\mathbf{z}'}[(\hat{\theta}_{\mathbf{x},\mathbf{z}} - \theta_0)'\nabla_\theta \log p(\mathbf{x}',\mathbf{z}';\theta_0)] = \mathbb{E}_{\mathbf{x},\mathbf{z}}[(\hat{\theta}_{\mathbf{x},\mathbf{z}} - \theta_0)']\mathbb{E}_{\mathbf{x}',\mathbf{z}'}[\nabla_\theta \log p(\mathbf{x}',\mathbf{z}';\theta_0)]$.

Moreover $\mathbb{E}_{\mathbf{x}',\mathbf{z}'}[\nabla_\theta \log p(\mathbf{x}',\mathbf{z}';\theta_0)] = 0$ owing to the definition of θ_0. Then, by the law of large numbers, we get $\frac{1}{n}\nabla^2_\theta \log p(\mathbf{x}',\mathbf{z}';\bar{\theta}) \xrightarrow{P} -J_\beta$. Using also that $\sqrt{n}(\hat{\theta}_{\mathbf{x},\mathbf{z}} - \theta_0) \xrightarrow{\mathcal{D}} \mathcal{N}(0, J_\beta^{-1})$ we get $\sqrt{n}(\hat{\theta}_{\mathbf{x},\mathbf{z}} - \theta_0)\sqrt{n}(\hat{\theta}_{\mathbf{x},\mathbf{z}} - \theta_0)' \xrightarrow{\mathcal{D}} \mathcal{W}_\nu(J_\beta^{-1}, 1)$, which is the Wishart distribution with scale parameter J_β^{-1} and one degree of freedom. From Slustky Lemma [20] we deduce

$$\nabla^2_\theta \log p(\mathbf{x}',\mathbf{z}';\bar{\theta})(\hat{\theta}_{\mathbf{x},\mathbf{z}} - \theta_0)(\hat{\theta}_{\mathbf{x},\mathbf{z}} - \theta_0)' \xrightarrow{\mathcal{D}} -J_\beta \mathcal{W}_\nu(J_\beta^{-1}, 1), \tag{30}$$

from which, it follows that

$$\mathbb{E}_{\mathbf{x},\mathbf{z},\mathbf{x}',\mathbf{z}'}[\mathrm{tr}[\nabla^2_\theta \log p(\mathbf{x}',\mathbf{z}';\bar{\theta})(\hat{\theta}_{\mathbf{x},\mathbf{z}} - \theta_0)(\hat{\theta}_{\mathbf{x},\mathbf{z}} - \theta_0)']] = -\nu + o(1) \tag{31}$$

and Equation (22) is proved.

We now prove Equation (23). A Taylor expansion of $\log p(\mathbf{x},\mathbf{z};\theta_0)$ about $\hat{\theta}_{\mathbf{x},\mathbf{z}}$ gives

$$\begin{aligned} 2\log p(\mathbf{x},\mathbf{z};\theta_0) &= 2\log p(\mathbf{x},\mathbf{z};\hat{\theta}_{\mathbf{x},\mathbf{z}}) - 2(\hat{\theta}_{\mathbf{x},\mathbf{z}} - \theta_0)'\nabla_\theta \log p(\mathbf{x},\mathbf{z};\hat{\theta}_{\mathbf{x},\mathbf{z}}) && (32)\\ &\quad +(\hat{\theta}_{\mathbf{x},\mathbf{z}} - \theta_0)'\nabla^2_\theta \log p(\mathbf{x},\mathbf{z};\bar{\theta})(\hat{\theta}_{\mathbf{x},\mathbf{z}} - \theta_0) && (33)\end{aligned}$$

where $\bar{\theta}$ is on the line joining θ_0 and $\hat{\theta}_{\mathbf{x},\mathbf{z}}$. Then, taking the expectation and noting the fact that $\nabla_\theta \log p(\mathbf{x},\mathbf{z};\hat{\theta}_{\mathbf{x},\mathbf{z}}) = 0$, leads to

$$\begin{aligned} 2\mathbb{E}_{\mathbf{x},\mathbf{z}}[\log p(\mathbf{x},\mathbf{z};\theta_0)] &= 2\mathbb{E}_{\mathbf{x},\mathbf{z}}[\log p(\mathbf{x},\mathbf{z};\hat{\theta}_{\mathbf{x},\mathbf{z}})] && (34)\\ &\quad +\mathbb{E}_{\mathbf{x},\mathbf{z}}[\mathrm{tr}[\nabla^2_\theta \log p(\mathbf{x},\mathbf{z};\bar{\theta})(\hat{\theta}_{\mathbf{x},\mathbf{z}} - \theta_0)(\hat{\theta}_{\mathbf{x},\mathbf{z}} - \theta_0)']] && (35)\end{aligned}$$

Finally, from the arguments used to prove Equation (31),

$$\mathbb{E}_{\mathbf{x},\mathbf{z}}[\mathrm{tr}[\nabla^2_\theta \log p(\mathbf{x},\mathbf{z};\bar{\theta})(\hat{\theta}_{\mathbf{x},\mathbf{z}} - \theta_0)(\hat{\theta}_{\mathbf{x},\mathbf{z}} - \theta_0)']] = -\nu + o(1), \tag{36}$$

which concludes the proof of Equation (23).

We now prove Equation (24). A second order Taylor expansion about θ_0 gives

$$\begin{aligned} 2\log p(\mathbf{x}';\hat{\theta}_{\mathbf{x},\mathbf{z}}) &= 2\log p(\mathbf{x}';\theta_0) + 2(\hat{\theta}_{\mathbf{x},\mathbf{z}} - \theta_0)'\nabla_\theta \log p(\mathbf{x}';\theta_0) && (37)\\ &\quad +(\hat{\theta}_{\mathbf{x},\mathbf{z}} - \theta_0)'\nabla^2_\theta \log p(\mathbf{x}';\bar{\theta})(\hat{\theta}_{\mathbf{x},\mathbf{z}} - \theta_0) && (38)\\ &= 2\log p(\mathbf{x};\theta_0) + 2(\hat{\theta}_{\mathbf{x},\mathbf{z}} - \theta_0)'\nabla_\theta \log p(\mathbf{x}';\theta_0) && (39)\\ &\quad +\mathrm{tr}[\nabla^2_\theta \log p(\mathbf{x}';\bar{\theta})(\hat{\theta}_{\mathbf{x},\mathbf{z}} - \theta_0)(\hat{\theta}_{\mathbf{x},\mathbf{z}} - \theta_0)'] && (40)\end{aligned}$$

with $\bar{\theta}$ on the line joining θ_0 and $\hat{\theta}_{\mathbf{x},\mathbf{z}}$. Then taking the expectation , we obtain

$$\begin{aligned} 2\mathbb{E}_{\mathbf{x},\mathbf{z},\mathbf{x}'}[\log p(\mathbf{x}';\hat{\theta}_{\mathbf{x},\mathbf{z}})] &= 2\mathbb{E}_{\mathbf{x},\mathbf{z},\mathbf{x}'}[\log p(\mathbf{x}';\theta_0)] + 2\mathbb{E}_{\mathbf{x},\mathbf{z},\mathbf{x}'}[(\hat{\theta}_{\mathbf{x},\mathbf{z}} - \theta_0)'\nabla_\theta \log p(\mathbf{x}';\theta_0)]\\ &\quad +\mathbb{E}_{\mathbf{x},\mathbf{z},\mathbf{x}'}[(\hat{\theta}_{\mathbf{x},\mathbf{z}} - \theta_0)'\nabla^2_\theta \log p(\mathbf{x}';\bar{\theta})(\hat{\theta}_{\mathbf{x},\mathbf{z}} - \theta_0)]\\ &= 2\mathbb{E}_{\mathbf{x},\mathbf{z},\mathbf{x}'}[\log p(\mathbf{x}';\theta_0)] + 2\mathbb{E}_{\mathbf{x},\mathbf{z},\mathbf{x}'}[(\hat{\theta}_{\mathbf{x},\mathbf{z}} - \theta_0)'\nabla_\theta \log p(\mathbf{x}';\theta_0)]\\ &\quad +\mathbb{E}_{\mathbf{x},\mathbf{z},\mathbf{x}'}[\mathrm{tr}(\nabla^2_\theta \log p(\mathbf{x}';\bar{\theta})(\hat{\theta}_{\mathbf{x},\mathbf{z}} - \theta_0)(\hat{\theta}_{\mathbf{x},\mathbf{z}} - \theta_0)')].\end{aligned}$$

The first term is given by $\mathbb{E}_{\mathbf{x},\mathbf{z},\mathbf{x}'}[\log p(\mathbf{x}';\theta_0)] = \mathbb{E}_{\mathbf{x}}[\log p(\mathbf{x};\theta_0)]$ since $\mathbf{x}$ and $\mathbf{x}'$ have the same distribution. The second term vanishes since firstly $\mathbb{E}_{\mathbf{x},\mathbf{z},\mathbf{x}'}[(\hat{\theta}_{\mathbf{x},\mathbf{z}} - \theta_0)'\nabla_\theta \log p(\mathbf{x}';\theta_0)] = \mathbb{E}_{\mathbf{x},\mathbf{z}}[(\hat{\theta}_{\mathbf{x},\mathbf{z}} - \theta_0)']\mathbb{E}_{\mathbf{x}'}[\nabla_\theta \log p(\mathbf{x}';\theta_0)]$ because of $\mathbf{x},\mathbf{z}$ and

$\mathbf{x}'$ independence, and secondly $\mathbb{E}_{\mathbf{x}'}[\nabla_\theta \log p(\mathbf{x}';\theta_0)] = 0$. For the third term, using the law of large numbers we get $\frac{1}{n}\nabla^2_\theta \log p(\mathbf{x}';\bar{\theta}) \xrightarrow{P} -J$, and from Lemma 1, we have

$$\sqrt{n}(\hat{\theta}_{\mathbf{x},\mathbf{z}} - \theta_0)\sqrt{n}(\hat{\theta}_{\mathbf{x},\mathbf{z}} - \theta_0)' \xrightarrow{\mathcal{D}} \mathcal{W}_\nu(J_\beta^{-1}, 1), \tag{41}$$

thus we deduce that

$$\nabla^2_\theta \log p(\mathbf{x}';\theta_0)(\hat{\theta}_{\mathbf{x},\mathbf{z}} - \theta_0)(\hat{\theta}_{\mathbf{x},\mathbf{z}} - \theta_0)' \xrightarrow{\mathcal{D}} -J\mathcal{W}_\nu(J_\beta^{-1}, 1). \tag{42}$$

Consequently

$$\mathbb{E}_{\mathbf{x},\mathbf{z},\mathbf{x}',\mathbf{z}'}[\mathrm{tr}[\nabla^2_\theta \log p(\mathbf{x}';\bar{\theta})(\hat{\theta}_{\mathbf{x},\mathbf{z}} - \theta_0)(\hat{\theta}_{\mathbf{x},\mathbf{z}} - \theta_0)']] = -\mathrm{tr}(JJ_\beta^{-1}) + o(1), \tag{43}$$

which concludes the proof of Equation (24).

The proof of Equation (25) is the same as for Equation (23) since $\hat{\theta}_{\mathbf{x}} \xrightarrow{P} \theta_0$.

To conclude the proof of Proposition 1, arguing the likelihoods $p(\mathbf{x},\mathbf{z};\hat{\theta}_{\mathbf{x},\mathbf{z}})$ and $p(\mathbf{x};\hat{\theta}_{\mathbf{x}})$ are concentrated about their maximums and decline fast as one moves away, then when the sample sizes are large enough, we replace the expectations at the right hand of both (23) and (25) by the observed values. This approximation introduces an error which by the central limit theorem is of order $O_p(\sqrt{n})$.

B Proof of Proposition 2

We introduce two lemmas (Lemmas 2 and 3) before proving Proposition 2.

Lemma 2 *Let* $\ell(\theta_1,\theta_2) = \log p(\mathbf{x},\mathbf{z};\theta_1) + \log p(\mathbf{x};\theta_2)$. *Then*

$$\frac{1}{\sqrt{n}}\nabla\ell(\theta_0,\theta_0) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \begin{pmatrix} J_\beta & J \\ J & J \end{pmatrix}\right). \tag{44}$$

Proof of Lemma 2

We have

$$\nabla\ell(\theta_0,\theta_0) = \sum_{i=1}^{n_\ell} \begin{pmatrix} \nabla_\theta \log p(\mathbf{x}_i,\mathbf{z}_i;\theta_0) \\ \nabla_\theta \log p(\mathbf{x}_i;\theta_0) \end{pmatrix} + \sum_{i=n_\ell+1}^{n} \begin{pmatrix} \nabla_\theta \log p(\mathbf{x}_i;\theta_0) \\ \nabla_\theta \log p(\mathbf{x}_i;\theta_0) \end{pmatrix}. \tag{45}$$

The key point is to apply a central limit theorem on labeled data and on unlabeled data, to get relation (44).

We note $\nabla_{\theta(i)} g(\theta_0)$ the derivate of $g(\theta)$ according to the ith component of θ, evaluated at θ_0. Centers are easily derived from $\mathbb{E}[\nabla_\theta \log p(\boldsymbol{x},z;\theta_0)] = \mathbb{E}[\nabla_\theta \log p(\boldsymbol{x};\theta_0)] = 0$. We now need to derive the variance matrices. For an unlabeled object, we have

$$\mathbb{E}[\nabla_{\theta(i)} \log p(\boldsymbol{x};\theta_0)\nabla_{\theta(j)} \log p(\boldsymbol{x};\theta_0)] = J_{ij}, \tag{46}$$

whereas for a labeled object we have

$$\mathbb{E}[\nabla_{\theta(i)} \log p(\boldsymbol{x},z;\theta_0)\nabla_{\theta(j)} \log p(\boldsymbol{x};\theta_0)] = \mathbb{E}_{\boldsymbol{x}}[\mathbb{E}_{z|\boldsymbol{x}}[\nabla_{\theta(i)} \log p(\boldsymbol{x},z;\theta_0)]\nabla_{\theta(j)} \log p(\boldsymbol{x};\theta_0)]. \tag{47}$$

Since

$$\mathbb{E}_{z|\boldsymbol{x}}[\nabla_{\theta(i)} \log p(\boldsymbol{x}, z; \theta_0)] = \sum_{k=1}^{g} \frac{\nabla_{\theta(i)} p(\boldsymbol{x}, k; \theta_0)}{p(\boldsymbol{x}, k; \theta_0)} p(k|\boldsymbol{x}) \tag{48}$$

and $p(z = k|\boldsymbol{x}) = p(z = k|\boldsymbol{x}; \theta_0)$, we have

$$\sum_{k=1}^{g} \frac{\nabla_{\theta(i)} p(\boldsymbol{x}, k; \theta_0)}{p(\boldsymbol{x}, k; \theta_0)} p(k|\boldsymbol{x}) = \sum_{k=1}^{g} \frac{\nabla_{\theta(i)} p(\boldsymbol{x}, k; \theta_0)}{p(\boldsymbol{x}; \theta_0)} = \frac{\nabla_{\theta(i)} p(\boldsymbol{x}; \theta_0)}{p(\boldsymbol{x}; \theta_0)} = \nabla_{\theta(i)} \log p(\boldsymbol{x}; \theta_0), \tag{49}$$

and thus

$$\mathbb{E}[\nabla_{\theta(i)} \log p(\boldsymbol{x}, z; \theta_0) \nabla_{\theta(j)} \log p(\boldsymbol{x}; \theta_0)] = \mathbb{E}[\nabla_{\theta(i)} \log p(\boldsymbol{x}; \theta_0) \nabla_{\theta(j)} \log p(\boldsymbol{x}; \theta_0)] = J_{ij}. \tag{50}$$

From the central limit theorem, we have

$$V_n = \sqrt{n_\ell} \left(\frac{1}{n_\ell} \sum_{i=1}^{n_\ell} \begin{pmatrix} \nabla_\theta \log p(\mathbf{x}_i, \mathbf{z}_i; \theta_0) \\ \nabla_\theta \log p(\mathbf{x}_i; \theta_0) \end{pmatrix} - 0 \right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \begin{pmatrix} J_c & J \\ J & J \end{pmatrix}\right), \tag{51}$$

and

$$U_n = \sqrt{n - n_\ell} \left(\frac{1}{n - n_\ell} \sum_{i=n_\ell+1}^{n} \begin{pmatrix} \nabla_\theta \log p(\mathbf{x}_i; \theta_0) \\ \nabla_\theta \log p(\mathbf{x}_i; \theta_0) \end{pmatrix} - 0 \right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \begin{pmatrix} J & J \\ J & J \end{pmatrix}\right). \tag{52}$$

Then using the independence of labeled and unlabeled data we get the result:

$$\frac{1}{\sqrt{n}} \nabla \ell(\theta_0, \theta_0) = \sqrt{\frac{n_\ell}{n}} V_n + \sqrt{\frac{n - n_\ell}{n}} U_n \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \begin{pmatrix} J_\beta & J \\ J & J \end{pmatrix}\right). \tag{53}$$

This result is useful for proving Lemma 3 below.

Lemma 3 $\sqrt{n}(\hat{\theta}_{\mathbf{x},\mathbf{z}} - \hat{\theta}_{\mathbf{x}}) \xrightarrow{\mathcal{D}} \mathcal{N}(0, J^{-1} - J_\beta^{-1})$.

Proof of Lemma 3

We have $\nabla \ell(\hat{\theta}_{\mathbf{x},\mathbf{z}}, \hat{\theta}_{\mathbf{x}}) = 0$. A Taylor expansion about θ_0 gives

$$\nabla \ell(\hat{\theta}_{\mathbf{x},\mathbf{z}}, \hat{\theta}_{\mathbf{x}}) = \nabla \ell(\theta_0, \theta_0) + \nabla^2 \ell(\bar{\theta}_1, \bar{\theta}_2) \begin{pmatrix} \hat{\theta}_{\mathbf{x},\mathbf{z}} - \theta_0 \\ \hat{\theta}_{\mathbf{x}} - \theta_0 \end{pmatrix} = 0 \tag{54}$$

with $(\bar{\theta}_1, \bar{\theta}_2)$ on the line joining $(\hat{\theta}_{\mathbf{x},\mathbf{z}}, \hat{\theta}_{\mathbf{x}})$ and (θ_0, θ_0), thus

$$\sqrt{n} \begin{pmatrix} \hat{\theta}_{\mathbf{x},\mathbf{z}} - \theta_0 \\ \hat{\theta}_{\mathbf{x}} - \theta_0 \end{pmatrix} = \left[-\frac{1}{n} \nabla^2 \ell(\bar{\theta}_1, \bar{\theta}_2) \right]^{-1} \frac{1}{\sqrt{n}} \nabla \ell(\theta_0, \theta_0). \tag{55}$$

Under standard regularity conditions, using the law of large numbers and the convergence $(\hat{\theta}_{\mathbf{x},\mathbf{z}}, \hat{\theta}_{\mathbf{x}}) \xrightarrow{P} (\theta_0, \theta_0)$, we establish that

$$\left[-\frac{1}{n} \nabla^2 \ell(\bar{\theta}_1, \bar{\theta}_2) \right]^{-1} \xrightarrow{P} \begin{pmatrix} J_\beta^{-1} & 0 \\ 0 & J^{-1} \end{pmatrix}. \tag{56}$$

Moreover from Lemma 2 we get

$$\sqrt{n}\begin{pmatrix}\hat{\theta}_{\mathbf{x},\mathbf{z}}-\theta_0\\ \hat{\theta}_{\mathbf{x}}-\theta_0\end{pmatrix}\xrightarrow{\mathcal{D}}\mathcal{N}\left(0,\begin{pmatrix}J_\beta^{-1} & J_\beta^{-1}\\ J_\beta^{-1} & J^{-1}\end{pmatrix}\right), \tag{57}$$

and finally we conclude the proof by using the continuous mapping theorem

$$\sqrt{n}(\hat{\theta}_{\mathbf{x},\mathbf{z}}-\hat{\theta}_{\mathbf{x}})\xrightarrow{\mathcal{D}}\mathcal{N}(0,J^{-1}-J_\beta^{-1}). \tag{58}$$

Proof of Proposition 2

We can now prove Proposition 2. From a second order Taylor expansion about $\hat{\theta}_{\mathbf{x}}$,

$$\begin{aligned}2\mathbb{E}_{\mathbf{x},\mathbf{z}}[\log p(\mathbf{x};\hat{\theta}_{\mathbf{x}})-\log p(\mathbf{x};\hat{\theta}_{\mathbf{x},\mathbf{z}})] &= -\mathbb{E}_{\mathbf{x},\mathbf{z}}[(\hat{\theta}_{\mathbf{x},\mathbf{z}}-\hat{\theta}_{\mathbf{x}})'\nabla^2\log p(\mathbf{x};\hat{\theta}_{\mathbf{x}})(\hat{\theta}_{\mathbf{x},\mathbf{z}}-\hat{\theta}_{\mathbf{x}})]+o(1)\\ &= -\mathbb{E}_{\mathbf{x},\mathbf{z}}[\mathrm{tr}(\sqrt{n}(\hat{\theta}_{\mathbf{x},\mathbf{z}}-\hat{\theta}_{\mathbf{x}})\sqrt{n}(\hat{\theta}_{\mathbf{x},\mathbf{z}}-\hat{\theta}_{\mathbf{x}})'\frac{1}{n}\nabla^2\log p(\mathbf{x};\hat{\theta}_{\mathbf{x}}))]+o(1).\end{aligned}$$

Then under standard regularity conditions $-\frac{1}{n}\nabla^2\log p(\mathbf{x};\hat{\theta}_{\mathbf{x}})\xrightarrow{P}J$ and using Lemma 3 we get

$$2\mathbb{E}_{\mathbf{x},\mathbf{z}}[\log p(\mathbf{x};\hat{\theta}_{\mathbf{x}})-\log p(\mathbf{x};\hat{\theta}_{\mathbf{x},\mathbf{z}})]=\mathrm{tr}(I-JJ_\beta^{-1})+o(1)=\nu-\mathrm{tr}(JJ_\beta^{-1})+o(1) \tag{59}$$

which concludes the proof after replacing the expectation with the observed value.

C Proof of Proposition 3

We need the following lemma for proving Proposition 3.

Lemma 4 *Let A and B two real $(d+p)$ symmetric positive definite matrices. If $A=\begin{pmatrix}a_{11} & a_{12}\\ a_{12}' & a_{22}\end{pmatrix}$ and $B=\begin{pmatrix}b_{11} & b_{12}\\ b_{12}' & b_{22}\end{pmatrix}$ where a_{11} and b_{11} are p symmetric matrices, a_{12} and b_{12} are in $\mathbb{R}^{p\times d}$ and a_{22} and b_{22} are d symmetric matrices, then $tr(BA^{-1})>tr(b_{22}a_{22}^{-1})$.*

Proof of Lemma 4

For the sake of simplicity we will prove the lemma for $p=1$. It can be shown that as A is symmetric positive definite, a_{22} and $a_{11}-a_{12}a_{22}^{-1}a_{12}'$ are symmetric definite positive matrices and then the inverse of A can be written

$$A^{-1}=\begin{pmatrix}c & cf'\\ cf & a_{22}^{-1}+cff'\end{pmatrix}=\begin{pmatrix}0 & 0\\ 0 & a_{22}^{-1}\end{pmatrix}+c\begin{pmatrix}1\\ f\end{pmatrix}\begin{pmatrix}1 & f'\end{pmatrix}$$

where $c=(a_{11}-a_{12}a_{22}^{-1}a_{12}')^{-1}\in\mathbb{R}$ and $f=-a_{22}^{-1}a_{12}'$. We have

$$BA^{-1}=\begin{pmatrix}0 & b_{12}a_{22}^{-1}\\ 0 & b_{22}a_{22}^{-1}\end{pmatrix}+cB\begin{pmatrix}1\\ f\end{pmatrix}\begin{pmatrix}1 & f'\end{pmatrix}$$

$$\mathrm{tr}(BA^{-1})=\mathrm{tr}(b_{22}a_{22}^{-1})+c\begin{pmatrix}1 & f'\end{pmatrix}B\begin{pmatrix}1\\ f\end{pmatrix}.$$

As B is definite positive, $\begin{pmatrix}1 & f'\end{pmatrix} B \begin{pmatrix}1\\ f\end{pmatrix} > 0$ which concludes the proof of the lemma for $p = 1$. Then for a general p we just need to apply the above result p times.

Proof of Proposition 3

We now can prove Proposition 3. Assume the data distribution belongs to two nested model m and m' with $m \subset m'$. Using Equation (59) in the proof of Proposition 2, we have

$$\begin{aligned}\mathbb{E}[\mathrm{AIC}_{cond}(m)] - \mathbb{E}[\mathrm{AIC}_{cond}(m')] &= 2\mathbb{E}[\mathrm{BEC}(m) - \mathrm{BEC}(m')] - [\nu_m - \mathrm{tr}(J(m)J_\beta(m)^{-1})] \\ &\quad + [\nu_{m'} - \mathrm{tr}(J(m')J_\beta(m')^{-1})] + o(1).\end{aligned}$$

Since $\mathbb{E}[\mathrm{BEC}(m) - \mathrm{BEC}(m')] \to 0$ (cf. [4], Proposition 2), we just need to prove that $-[\nu_m - \mathrm{tr}(J(m)J_\beta(m)^{-1})] + [\nu_{m'} - \mathrm{tr}(J(m')J_\beta(m')^{-1})] > 0$. Writing

$$\begin{aligned}&-[\nu_m - \mathrm{tr}(J(m)J_\beta(m)^{-1})] + [\nu_{m'} - \mathrm{tr}(J(m')J_\beta(m')^{-1})] = \\ &\quad \mathrm{tr}[(J_\beta(m') - J(m'))J_\beta(m')^{-1}] - \mathrm{tr}[(J_\beta(m) - J(m))J_\beta(m)^{-1}],\end{aligned}$$

we make the decompositions $J(m') = \begin{pmatrix}J^{11}(m') & J^{12}(m')\\ J^{21}(m') & J(m)\end{pmatrix}$ and $J_\beta(m') = \begin{pmatrix}J_\beta^{11}(m') & J_\beta^{12}(m')\\ J_\beta^{21}(m') & J_\beta(m)\end{pmatrix}$. It is important to notice that $J_\beta(m) - J(m)$ and $J_\beta(m') - J(m')$ are positive definite matrices.

Taking $A = J_\beta(m')$, $B = J_\beta(m') - J(m')$, $a_{22} = J_\beta(m)$, $b_{22} = J_\beta(m) - J(m)$, and applying Lemma 4, we conclude that $-[\nu(m) - \mathrm{tr}(J(m)J_\beta(m)^{-1})] + [\nu(m') - \mathrm{tr}(J(m')J_\beta(m')^{-1})] > 0$, and consequently that for n large enough

$$\mathbb{E}[\mathrm{AIC}_{cond}(m)] - \mathbb{E}[\mathrm{AIC}_{cond}(m')] > 0 \tag{60}$$

which concludes the proof.

References

[1] H. Akaike. Information theory and an extension of the maximum likelihood principle. *Second International Symposium on Information Theory*, pages 267–281, 1974.

[2] J. A. Anderson and S. C. Richardson. Logistic discrimination and bias correction in maximum likelihood estimation. *Technometrics*, 21(1):71–78, 1979.

[3] C. Biernacki, F. Beninel, and V. Bretagnolle. A generalized discriminant rule when training population and test population differ on their descriptive parameters. *Biometrics*, 58(2):387–397, 2002.

[4] G. Bouchard and G. Celeux. Selection of generative models in classification. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 28(4):544–554, 2006.

[5] G. Celeux and G. Govaert. Parsimonious Gaussian models in cluster analysis. *Pattern Recognition*, 28:781–793, 1995.

[6] O. Chapelle, B. Schölkopf, and A. Zien, editors. *Semi-Supervised Learning*. MIT Press, Cambridge, MA, 2006.

[7] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[8] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188, 1936.

[9] C. Fraley and A. E. Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97:611–631, 2002.

[10] A. Goldenshluger and E. Greenshtein. Asymptotically minimax regret procedures in regression model selection and the magnitude of the dimension penalty. *The Annals of Statistics*, 28:1620–1637, 2000.

[11] T. Hastie and R. Tibshirani. Discriminant analysis by Gaussian mixtures. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58:155–176, 1996.

[12] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning (second edition)*. Springer Series in Statistics, 2009.

[13] G. Mclachlan. *Discriminant Analysis and Statistical Pattern Recognition*. Wiley Series in Probability and Statistics. Wiley-Interscience, 2004.

[14] G. McLachlan and D. Peel. *Finite mixutre models*. Wiley Series in Probability and Statistics: Apllied Probability and Statistics, Wiley-Interscience, New York, 2000.

[15] D. J. Miller and J. Browning. A mixture model and EM-based algorithm for class discovery, robust classification, and outlier rejection in mixed labeled/unlabeled data sets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(11):1468–1483, 2003.

[16] T. O'Neill. The general distribution of the error rate of a classification procedure with application to logistic regression discrimination. *Journal of the American Statistical Association*, 75(369):154–160, 1980.

[17] F. Rosenblat. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65:386–408, 1958.

[18] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

[19] M. Stone. An asymptotic equivalence of choice of model by cross-validation and akaike's criterion. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1977.

[20] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, June 2000.

[21] V. Vandewalle. *Estimation et sélection en classification semi-supervisée*. PhD thesis, Université de Lille 1, 2009.

[22] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, 1995.

Centre de recherche INRIA Saclay – Île-de-France
Parc Orsay Université - ZAC des Vignes
4, rue Jacques Monod - 91893 Orsay Cedex (France)

Centre de recherche INRIA Bordeaux – Sud Ouest : Domaine Universitaire - 351, cours de la Libération - 33405 Talence Cedex
Centre de recherche INRIA Grenoble – Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier
Centre de recherche INRIA Lille – Nord Europe : Parc Scientifique de la Haute Borne - 40, avenue Halley - 59650 Villeneuve d'Ascq
Centre de recherche INRIA Nancy – Grand Est : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex
Centre de recherche INRIA Paris – Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex
Centre de recherche INRIA Rennes – Bretagne Atlantique : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex
Centre de recherche INRIA Sophia Antipolis – Méditerranée : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399