Journal of Machine Learning Research 15 (2014) 367-443 Submitted 1/13; Revised 9/13; Published 2/14

Unbiased Generative Semi-Supervised Learning

Patrick Fox-Roberts[*] PFOXROBERTS@GMAIL.COM
Cambridge University Engineering Department
Trumpington Street
Cambridge, CB2 1PZ, UK

Edward Rosten ED@COMPUTERVISIONCONSULTING.COM
Computer Vision Consulting
7th floor
14 Bonhill Street
London, EC2A 4BX, UK

Editor: William Cohen

Abstract

Reliable semi-supervised learning, where a small amount of labelled data is complemented by a large body of unlabelled data, has been a long-standing goal of the machine learning community. However, while it seems intuitively obvious that unlabelled data can aid the learning process, in practise its performance has often been disappointing. We investigate this by examining generative maximum likelihood semi-supervised learning and derive novel upper and lower bounds on the degree of bias introduced by the unlabelled data. These bounds improve upon those provided in previous work, and are specifically applicable to the challenging case where the model is unable to exactly fit to the underlying distribution a situation which is common in practise, but for which fewer guarantees of semi-supervised performance have been found. Inspired by this new framework for analysing bounds, we propose a new, simple reweighing scheme which provides a provably unbiased estimator for arbitrary model/distribution pairs—an unusual property for a semi-supervised algorithm. This reweighing introduces no additional computational complexity and can be applied to very many models. Additionally, we provide specific conditions demonstrating the circumstance under which the unlabelled data will lower the estimator variance, thereby improving convergence.

Keywords: Kullback-Leibler, semi-supervised, asymptotic bounds, bias, generative model

1. Introduction

Reliable semi-supervised learning has been a long standing goal of the machine learning community. Its desirability is motivated by the observation that when collecting data sets, often each sample has two distinct parts: some feature X, collected from some real world population, often consisting of one or more basic measurements; and some label Y, assigned by the experimenter, representing a higher level concept. Furthermore the act of assigning this higher level label very often constitutes a major bottleneck in the data set creation process. It is perhaps expensive (requiring an expert's opinion), slow (requiring

[*] PFR is currently affiliated with The Randal Division, Guy's Campus, King's College London, SE1 1UL

an investment of time or staff), or in some way destructive (requiring a component to be tested to destruction, or the death of a patient).

If we wish to fit a model parametrised by some set of parameters θ to this distribution, we will need some data set D_L consisting of N_L labelled samples, $D_L = (x_i, y_i)_{i=1,...,N_L}$, to train our model with; for example, if we are training using maximum likelihood, we must find the parameters which maximise $P(D_L|\theta)$, which for iid data is equivalent to finding

$$\theta^\star = \arg\max_\theta \sum_{i=1}^{N_L} \log\left(P(x_i, y_i|\theta)\right). \tag{1}$$

In order to get a solution which generalises well to unseen data N_L may have to be quite large, especially if the model is rich or the feature space $\mathcal{X}$ high dimensional.

A far preferable situation would be to be able to utilise a smaller labelled data set D_L, augmented with an additional data set D_U consisting of N_U unlabelled samples, $D_U = (x_i)_{i=N_L+1,...,N_L+N_U}$, which consist only of their observed feature rather than a feature - label pair. In essence the unlabelled data is used to 'bootstrap' the labelled. Unlabelled samples tell us the shape of our distribution in the feature space, while labelled samples give us the classification information.

At first glance, utilising unlabelled data to aid in fitting the parameters of some model appears trivial. Inspired by the likelihood principal (Jaynes, 2003), it is tempting to simply augment the likelihood function of the parameters, $P(D_L|\theta)$, with the additional unlabelled data, $P(D_L, D_U|\theta)$, and proceed with training exactly as before, that is, find the parameters

$$\theta_S^\star = \arg\max_\theta \sum_{i=1}^{N_L} \log\left(P(x_i, y_i|\theta)\right) + \sum_{i=N_L}^{N_L+N_U} \log\left(P(x_i|\theta)\right). \tag{2}$$

In practice however this has proven to give mixed results, sometime improving model fitting, other times worsening it. This unpredictable of performance has formed a very major barrier to more widespread adoption of semi-supervised techniques. Many alternative algorithms have been developed to counter this. However, there still exists a need to better understand and quantify why more standard methods fail.

This paper examines the effect of including unlabelled data in a training set when performing maximum likelihood fitting of generative models. In particular, it is well known (see, for example, Bishop, 2006) that maximising the parameter likelihood for labelled data approximately minimises the Kullback Leibler divergence between the parametric distribution $P(X, Y|\theta)$ and the underlying distribution the data is sampled from, $P(X, Y)$. We show that maximising the likelihood of a data set containing unlabelled samples minimises a different divergence. We then show that the possible error between this and the correct divergence may grow rapidly with the proportion of unlabelled data, and will do so monotonically.

Out of necessity, the analysis presented shall only concern itself with generative models. This follows in the footsteps of numerous other pieces of work which have shed light on the generative semi-supervised learning problem, for example Castelli and Cover (1995), Castelli and Cover (1996), Dillon et al. (2010), Cozman et al. (2003) and Yang and Priebe (2011). Moreover, generative models are of interest in and of themselves. For example, they are used in the fields of computer vision and text analysis, both of which could potentially benefit from better semi-supervised algorithms; recent examples of such work include that of Rauschert and Collins (2012), Beecks et al. (2011), Lücke and Eggert (2010), Kang et al. (2012) and Zhuang et al. (2012). In the general case there is also evidence that generative models can converge faster than discriminative, as shown by Ng and Jordan (2002), and so are valuable when dealing with small data sets.

2. Previous Work

A great deal of work has been done proposing algorithms designed to take advantage of semi-supervised data. Here we shall concern ourselves instead with examining the work done on finding general bounds on performance.

We begin by considering the highly influential work by Castelli and Cover (1995, 1996). This looks not at a particular semi-supervised algorithm, but rather at a slightly more general question of when unlabelled samples can be of value. They conclude that for an identifiable (as defined in the paper) binary decision problem, using a generative model, the misclassification risk decreases exponentially fast towards the Bayes error as the number of labelled samples increases. This result is encouraging. However, the requirement of identifiability is a strict one. In practise it cannot often be guaranteed, and may even be flatly contradicted.

The work of Dillon et al. (2010) builds upon this. Amongst other things they confirm that provided a data set is generated from $P(X, Y|\theta_0)$ where $\theta_0 \in \Omega$, the estimator

$$\hat{\theta}_N = \arg\max_{\theta \in \Omega} \sum_{i=1}^{N_L} \log(P(x_i, y_i|\theta)) + \sum_{i=N_L+1}^{N} \log(P(x_i|\theta))$$

is consistent. As such, in cases where there is good reason to believe the true distribution is drawn from the same family as our parametric model, we can expect consistent convergence. They also provide one of few examinations of the associated variance of an estimator, though again under the assumption of an identifiable model.

In a similar vein Zhang (2000) examines the fisher information matrix when learning parameters for semi-supervised learning, and conclude that even when their true distribution can not be expressed by the model parameters being fitted, unlabelled samples always aid in learning in that they reduce the variance of the estimator. From this work we can conclude that adding unlabelled samples is not preventing consistent convergence. As such, if performance is observed to often worsen instead of improve as the number of unlabelled samples increases, the fault must lie elsewhere.

The asymptotic behaviours of semi-supervised learning where the model is mis-specified has been further studied by Cozman et al. (2003); Cozman and Cohen (2006, 2002), where no assumptions are made about the parametric model being close to the underlying distribution. In particular, they show that the limiting value of the optimum parameters $\theta^\star$ when performing ML semi-supervised learning in such a scenario is

$$\arg\max_{\theta} \big((1-\lambda) E_{P(X,Y)} (\log P(x, y|\theta)) + \lambda E_{P(x)} (\log(P(x|\theta)))\big)$$

where λ is the probability of a sample being unlabelled. If λ varies (say by adding unlabelled samples) then this will likely change the optimal parameters $\theta^\star$, and so the associated error rate. In the limit, as $\lambda \to 1$, we will tend towards the solution found training entirely on unlabelled data. They argue that with a few assumptions on the modelling densities, $\theta^\star$ is a continuous function of λ. They also show that an instance where the asymptotically optimal parameters are not changed by λ comes, as might be expected, when the model is "correct" and can be fitted exactly to the underlying distribution (i.e., the true distribution $P(x, y)$ is a member of the family of distributions that can be modelled by $P(x, y|\theta)$).

The relative value of labelled/unlabelled samples was also investigated in Ratsaby and Venkatesh (1995) for the case of classifying between two multivariate gaussian distributions of unknown class prior and position parameters. As in the work by Castelli and Cover, an exponential decrease in error rate with the number of labelled samples is shown, and an only polynomial decrease in the same with the number of unlabelled samples. However they also demonstrate a deleterious effect in the *dimensionality* of the space, indicating unlabelled samples are likely to be less useful in high dimensional spaces. Separately, the work of Shahshahani and Landgrebe (1994) examines learning the parameters of both a single gaussian and a GMM when labels are missing. They too note an interesting effect of dimensionality on semi-supervised learning, in particular from the point of view of the Hughes phenomenon (Hughes, 1968). This is the observation that, in theory, increasing the dimensionality of a classification problem by taking new measurements should never increase the Bayes error; yet in practise, if we are learning from sampled data we find performance will after a while degrade due to the larger number of parameters that must be estimated (this is very closely linked to the perhaps more familiar *Curse of Dimensionality*, see Bishop, 2006). They propose that semi-supervised learning can help mitigate this, but only if the rate of introduction of bias due to the unlabelled samples is lower than the decrease in variance of the estimator.

Recently, Yang and Priebe (2011) has provided an investigation of semi-supervised generative learning that builds upon these conclusions. The key parameters they identify are the asymptotic optima achieved when performing fully supervised learning, θ^*_{sup}, and those achieved from entirely un-supervised learning, θ^*_{unsup}. Provided that the ratio of N_L to N tends towards 0 as N tends towards infinity (where $N = N_L + N_U$) we have the scenario where we are moving from a high-variance, unbiased estimate, towards a low variance, biased estimate. Interestingly, the KL divergences between the distributions defined by θ^*_{sup} and θ^*_{unsup}, and between these distributions and a given estimate based on a data set (either fully labelled or a mixture of labelled and unlabelled), are identified as providing bounds on the probability that classification performance will improve/worsen as unlabelled data is added. Intuitively, if the divergence between the models specified by θ^*_{sup} and θ^*_{unsup} is small, then adding unlabelled data is less likely to significantly worsen results. They also show for a particular model that the point at which this occurs can be quite sharp. However, as it is likely to be different for different models and distributions, it still remains an open question how it can be best estimated.

Additionally, theoretic examinations of expected performance for other semi supervised learning situations, such as transductive learning (for example, Wang et al., 2007; Vapnik, 1998), PAC learning (Balcan and Blum, 2005; Blum and Balcan, 2010), and for generic loss functions (Syed and Taskar, 2010), have also been carried out. However, as our purpose here is to examine what can be said about generative semi-supervised learning we shall not discuss these further.

2.1 Non-ML Algorithms

Given the problems associated with standard ML semi-supervised learning, as well as the desire to utilise unlabelled samples in non-generative models, a large number of alternative objective functions have been proposed to take advantage of unlabelled data. Notable examples include Multi Conditional Learning (introduced by McCallum et al. 2006 and applied to semi-supervised learning by Druck et al. 2007) and the hybrid Bayesian approach of Lasserre et al. (2006), both of which utilise mixtures of generative and discriminative models; information theory based approaches, which consider the similarity of class predictions across the kNN graph such as Subramanya and Bilmes (2009, 2008), the mutual information of samples within local clusters (Szummer and Jaakkola, 2002), or the conditional entropy of class predictions across the unlabelled samples (Grandvalet and Bengio, 2006); Expectation Regularisation (Mann and McCallum, 2007), which seeks to enforce class proportion constraints; Co-training, (Blum and Mitchell, 1998), which makes use of situations where data is known to be separable in two different 'views'; transduction, (Vapnik, 1998), and the transductive support vector machine; kernel methods, such as those investigated by Krishnapuram et al. (2005) and Jaakkola and Haussler (1999), which seek to use unlabelled samples to build better kernel functions; and many others. A thorough literature review was carried out by Zhu (2005).

3. Local And Global Bounds On Semi-Supervised Divergences

We now present a number of theorems, showing the asymptotic limits of the performance of models trained on semi-supervised data using the standard technique Equation (2). While it has been previously noted in the literature that ML semi-supervised learning introduces bias when the model and underlying distributions do not match, we provide new bounds on the degree of this bias as a function of the proportion of unlabelled data, and the best case performance of our model if it were to be trained on a large labelled data set, giving new insight into the reason behind these bounds.

3.1 Notation And Conditions

A semi-supervised data set consists of two types of data - labelled samples drawn from $P(X,Y)$, and unlabelled drawn from $P(X)$. To allow us to deal with both of these within a single framework we shall introduce a new variable Z, and consider our *entire* data set to be drawn from $P(X,Z)$, $\{x_i, z_i\}_{i=1,\ldots N}$ in the space $\mathcal{X} \times \mathcal{Z}$. We shall allow the 'labelling' Z to take on the same set of values as Y, plus one extra, U, and therefore $\mathcal{Z} = \mathcal{Y} \cup U$. For every "labelled" sample, $z_i = y_i$, and for every "unlabelled" sample $z_i = U$. As such we now have the data set $\{x_i, z_i\}_{i=1,\ldots N}$. Similarly, we shall consider our parameters θ to specify a distribution $P(X,Z|\theta)$ rather than $P(X,Y|\theta)$, in a manner which will become clear as we proceed.

We shall now apply several conditions to $P(X,Z|\theta)$ and $P(X,Z)$ to allow them to reflect what we consider the typical maximum likelihood generative semi-supervised learning problem.

Condition 1 *X is conditionally independent of Z given Y — if we know the class y of sample x, z gives us no more information, that is,*

$$P(x|y, z, \theta) = P(x|y, \theta), \quad P(x|y, z) = P(x|y)$$

This first condition represents the fact that z_i can be considered a noisy estimate of y_i - in as much as it will either be equal to y_i, or it will take on the value U to indicate y_i is unknown. In either case, if we had access to the true value of y_i, then z_i would be irrelevant as it can give us no useful information. This condition is similar to the "missing at random" assumption discussed by Grandvalet and Bengio (2006).

Condition 2 *The labelled samples have been drawn randomly and labelled correctly. The unlabelled samples are similarly drawn randomly, with no class bias. As such,*

$$P(z|y,\theta) = \left\{ \begin{array}{rl} P(U|\theta), & z = U \\ \delta_k(z,y)P(\bar{U}|\theta), & z \neq U \end{array} \right. \qquad P(z|y) = \left\{ \begin{array}{rl} P(U), & z = U \\ \delta_k(z,y)P(\bar{U}), & z \neq U \end{array} \right.$$

where δ_k indicates the Kronecker delta function, and we have denoted $1 - P(U|\theta)$ as $P(\bar{U}|\theta)$ and $1 - P(U)$ as $P(\bar{U})$

This second condition specifies our labelling process. It is imagined that a 'bag' full of unlabelled samples initially exists, and individual ones are then drawn from it and the correct label associated with them by some expensive labelling process[1] to form the labelled set.

In practise truly drawing samples completely at random runs with risk of certain classes having zero labelled samples, which is likely to cause highly undesirable behaviour of the algorithm. We do not foresee this as a problem for two reasons. Firstly, in the asymptotic limit (which is what most of our work will be concerned with in this section) we will almost surely achieve labelled samples being drawn from all classes. Secondly, in practise, the individuals running the experiment are likely to ensure that all classes have some representative samples. This breaks the assumption of iid data; however, provided the class priors are respected when choosing how many samples to label from each class (or suitable weighting applied) we can still attain an asymptotically unbiased estimate of the expectation term in the divergence.

Condition 3 *The proportion of labelled data is known, letting us set $P(\bar{U}|\theta) = P(\bar{U}) = 1 - P(U)$*

We assume this as matching labels to samples is a process controlled entirely by the user, and that they use this knowledge to set $P(U|\theta)$ rather than having to infer it from the data.

[1] This can be considered a somewhat simpler model to that proposed by Rosset et al. (2005), where the labelling also depended on the feature vector x. This work however is interested in the case of biased semi-supervised learning, which we assume here to not be the case.

3.1.1 Divergences

The KL divergence is a widely used method of measuring the similarity between two distributions, and one which shall be made extensive use of in this article. For distributions $P(X, Y)$ and $P(X, Y|\theta)$ where X is a continuous random variable and Y is discrete, it is defined as

$$KL(P(X,Y)||P(X,Y|\theta)) = \int_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} P(x,y) \log\left(\frac{P(x,y)}{P(x,y|\theta)}\right)$$

It is perhaps most widely used as a justification of maximum likelihood methods, as it is a standard proof that the parameters

$$\theta^\star = \arg\max_\theta \sum_{i=1}^{N_L} \log\left(p\left(x_i, y_i|\theta\right)\right)$$

are an asymptotically unbiased minimiser of $KL(P(X,Y)||P(X,Y|\theta))$, for example, see Bishop (2006).

For brevity and to make subsequent equations more readable we shall introduce a more concise notation to refer to the KL divergence. For random variables A and B and parameters θ, the full divergence shall be denoted as

$$D(P(A,B),\theta) \equiv KL(P(A,B)||P(A,B|\theta))$$

and the conditional divergence as

$$D(P(A|B),\theta) \equiv KL(P(A|B)||P(A|B,\theta)).$$

3.2 Standard ML Semi-Supervised Learning Expressed As A Divergence

Our first step is to demonstrate that when a proportion of a data set whose likelihood we are maximising is lacking labels, in the asymptotic limit we will minimise a different divergence to that we might wish - specifically, we minimise $D(P(X,Z)|\theta)$ rather than $D(P(X,Y)|\theta)$.

Theorem 4 *Subject to the conditions in 3.1, maximising*

$$\prod_{i=1}^{N_L} P(x_i, y_i|\theta) \prod_{i=n_L+1}^{N_L+N_U} P(x_i|\theta) \tag{3}$$

w.r.t. θ minimises an asymptotically unbiased estimate of a term directly proportional to $D(P(X,Z),\theta)$, not $D(P(X,Y),\theta)$

Proof Given a set of N samples $\{x_i, z_i\}_{i=1,\ldots N}$ drawn from $P(X, Z)$, we can approximate the expectation term in $D(P(X,Z)|\theta)$ with an arithmetic mean over our samples (for example, see MacKay, 2003). Ignoring terms which are not a function of θ, and taking the antilog, we attain

$$\arg\min_\theta D(P(X,Z),\theta) \approx \arg\max_\theta \prod_{i=1}^{N} P(x_i, z_i|\theta). \tag{4}$$

Examining the above, and making use of Condition 1 to simplify $P(x_i|z_i, y, \theta)$ into $P(x_i|y, \theta)$, we can rewrite the likelihood contribution of a single sample i, $P(x_i, z_i|\theta)$ as follows

$$\begin{aligned} P(x_i, z_i|\theta) &= \sum_{y \in \mathcal{Y}} P(x_i, z_i, y|\theta) \\ &= \sum_{y \in \mathcal{Y}} P(x_i|z_i, y, \theta) P(z_i|y, \theta) P(y|\theta) \\ &= \sum_{y \in \mathcal{Y}} P(x_i|y, \theta) P(z_i|y, \theta) P(y|\theta) \\ &= \sum_{y \in \mathcal{Y}} P(x_i, y|\theta) P(z_i|y, \theta). \end{aligned}$$

This can be simplified further using Conditions 2 and 3, depending on the value of z_i. First consider the case where $z_i = U$

$$P(x_i, z_i|\theta)|_{z_i=U} = \sum_{y \in \mathcal{Y}} P(x_i, y|\theta) P(U|\theta) = P(x_i|\theta) P(U). \tag{5}$$

Thus, a sample whose labelling z_i indicates it is unlabelled contributes a quantity proportional to $P(x_i|\theta)$ to our likelihood expression. Now consider a single labelled example (i.e., where $z_i \neq U$),

$$\begin{aligned} P(x_i, z_i|\theta)|_{z_i \neq U} &= \sum_{y \in \mathcal{Y}} P(x_i, y|\theta) \delta_k(y, z_i) P(\bar{U}|\theta) \\ &= P(x_i, y|\theta)|_{y=z_i} P(\bar{U}) \\ &= P(x_i, y_i|\theta) P(\bar{U}) \end{aligned} \tag{6}$$

where we have made a slight change of notation in the last term to represent that if $z_i \neq U$, then y_i is known. This contributes a term proportional to $P(x_i, y_i|\theta)$ to the likelihood. If we substitute these results back into Equation (4), our final likelihood expression is

$$\arg\max_{\theta} \prod_{i, z_i \neq U} P(\bar{U}) P(x_i, y_i|\theta) \prod_{j, z_j = U} P(U) P(x_j|\theta)$$

which is equivalent to maximising Equation (3). ■

The form of our final likelihood function is the same as that found by Cozman and Cohen (2006). The difference is in the interpretation of this. While Cozman and Cohen (2006) considers it simply as a biased approximate minimiser of $D(P(X, Y), \theta)$, we consider it an *unbiased* minimiser of a *new divergence* $D(P(X, Z), \theta)$. The utility of this is that we can investigate the 'bias' introduced by considering the relationship between this semi-supervised divergence and the original fully supervised one, and the properties of KL divergences.

3.3 Bounding $D\big(P(X,Y),\theta\big)$ With $D\big(P(X,Z),\theta\big)$

Maximising the likelihood of a partially labelled data set corresponds to approximately minimising $D\left(P(X,Z),\theta\right)$. We now examine how $D\left(P(X,Z),\theta\right)$ is related to $D\left(P(X,Y),\theta\right)$, and show that a set of upper and lower bounds can be formed using it.

Theorem 5 *Subject to the conditions in 3.1, for a given set of parameters θ, $D\big(P(X,Z),\theta\big)$ defines an upper and lower set of bounds on $D\big(P(X,Y),\theta\big)$ as follows:*

$$D\big(P(X,Z),\theta\big) \leq D\big(P(X,Y),\theta\big) \leq \frac{D\big(P(X,Z),\theta\big)}{P(\bar{U})}. \tag{7}$$

Remark 6 *These bounds imply that, for a given $D\big(P(X,Z),\theta\big)$ that we are optimising, the divergence of interest $D\big(P(X,Y),\theta\big)$ could vary by up to a factor $P(\bar{U})^{-1}$. In situations where $P(\bar{U})^{-1}$ is large, this uncertainty may become the dominant factor in determining the quality of our result.*

Proof Consider the KL divergence $D(P(X,Z),\theta)$. We shall take the summation over Z and split out the term $z = U$, noting that $\mathcal{Z} - U = \mathcal{Y}$

$$\begin{aligned} D(P(X,Z),\theta) = &\int_{x\in\mathcal{X}} \sum_{z\in\mathcal{Y}} P(x,z)\log\Big(\frac{P(x,z)}{P(x,z|\theta)}\Big)dx \\ &+ \int_{x\in\mathcal{X}} P(x,z)|_{z=U}\log\Big(\frac{P(x,z)|_{z=U}}{P(x,z|\theta)|_{z=U}}\Big)dx. \end{aligned}$$

Using Equation (5) and Equation (6), and their corresponding counterparts when not conditioned on θ, we can simplify the terms within our logarithms

$$\frac{P(x,z)|_{z=U}}{P(x,z|\theta)|_{z=U}} = \frac{P(x)P(U)}{P(x|\theta)P(U)} = \frac{P(x)}{P(x|\theta)},$$

$$\frac{P(x,z)|_{z\neq U}}{P(x,z|\theta)|_{z\neq U}} = \frac{P(x,y)|_{y=z}}{P(x,y|\theta)|_{y=z}}.$$

Using these identities, and Equation (5) and Equation (6) this allows us to rewrite the divergence $D(P(X,Z),\theta)$

$$\begin{aligned} D(P(X,Z),\theta) = P(\bar{U}) &\int\limits_{x\in\mathcal{X}} \sum_{z\in\mathcal{Y}} P(x,y)|_{y=z}\log\Big(\frac{P(x,y)|_{y=z}}{P(x,y|\theta)|_{y=z}}\Big)dx \\ &+P(U)\int\limits_{x\in\mathcal{X}} P(x)\log\Big(\frac{P(x)}{P(x|\theta)}\Big)dx. \end{aligned}$$

which is exactly equivalent to

$$D(P(X,Z),\theta) = P(\bar{U})D\big(P(X,Y),\theta\big) + P(U)D\big(P(X),\theta\big). \tag{8}$$

From this we can find a set of upper and lower bounds on $D(P(X,Y),\theta)$ in terms of $D(P(X,Z),\theta)$ alone. The upper bound can be found by noting $D(P(X),\theta) \geq 0$, which given Equation (8) implies

$$D(P(X,Z),\theta) \geq P(\bar{U})D\big(P(X,Y),\theta\big) \tag{9}$$

which when rearranged gives the upper bound in Equation (7). The lower bound follows similarly, by noting that $D(P(X,Y),\theta) = D(P(Y|X),\theta) + D(P(X),\theta) \geq D(P(X),\theta)$. Again using Equation (8) this gives

$$\begin{aligned} D(P(X,Z),\theta) &\leq P(\bar{U})D\big(P(X,Y),\theta\big) + P(U)D\big(P(X,Y),\theta\big) \\ &= \big(P(\bar{U}) + P(U)\big)D\big(P(X,Y),\theta\big) \\ &= D\big(P(X,Y),\theta\big) \end{aligned} \tag{10}$$

which gives us our lower bound. Combining Equation (10) and Equation (9) gives us Equation (7). ■

It is notable that in deriving these bounds we have treated $D(P(X),\theta))$ (or, equivalently, $D(P(Y|X),\theta)))$ simply as a value in the range 0 to $D(P(X,Y),\theta)$. As we wished to find general bounds that would hold for any combination of $P(X,Y)$ and $P(X,Y|\theta)$ we feel that this is an entirely justifiable method of proceeding.

In practice, however, we will not be dealing with arbitrary distributions for $P(X,Y)$ and $P(X,Y|\theta)$; rather, $P(X,Y)$ will usually represent some measurements of a real world phenomenon that we believe to be learnable in (hopefully) some well chosen space. Similarly our model may have been selected from a pool of potential models are that which is considered most likely (according to some prior beliefs) to be able to fit to the distribution of interest acceptably well, and will also often be smoothly varying with non-negligible correlations between $P(Y|X,\theta)$ and $P(X|\theta)$. As such, with additional problem specific knowledge, we suspect that tighter bounds on $D(P(X,Y),\theta)$ will tend to exist.

Another question one might raise is whether the lower bound can become tight even in instances where there is a mismatch between the model and true distribution - that is, given $\min_\theta D(P(X,Y),\theta) > 0$, can we have the situation where $D(P(X,Z),\theta) = D(P(X,Y),\theta)$? To answer this, consider $D(P(X,Z),\theta)$ as written in Equation (8). This can be re-written as follows

$$D(P(X,Z),\theta) = D\big(P(X,Y),\theta\big) - P(U)D\big(P(Y|X),\theta\big).$$

As such, in order to achieve the situation where $D(P(X,Z),\theta) = D\big(P(X,Y),\theta\big)$, it must be the case that $P(U)D\big(P(Y|X),\theta\big) = 0$. Assuming $P(U) > 0$ (as otherwise we are dealing with the trivial case of utilising no labelled data) then this must mean $D(P(Y|X),\theta) = 0$, that is, the conditional distribution specified by the model perfectly matches the true distribution. This observation seems to match intuition - if the model can correctly predict the class of unlabelled data, then its divergence estimate will not be biased by utilising these samples.

3.4 Global Bounds

The results in 3.3 give us bounds on $D(P(X,Y),\theta)$ in terms of $D(P(X,Z),\theta)$ for a given θ. It is of more interest however to characterise the global minimisers of these two expressions.

That is, if we make use of our unlabelled data to minimise $D(P(X,Z),\theta)$ with respect to θ, what can be inferred about the value of $D(P(X,Y),\theta)$ evaluated at this minimum?

Theorem 7 *Define the optimum parameters for the supervised and ML semi-supervised learning problems as*

$$\theta^\star = \arg\min_\theta D(P(X,Y),\theta),$$
$$\theta_S^\star = \arg\min_\theta D(P(X,Z),\theta).$$

Subject to the conditions in 3.1, it can be shown that

$$D(P(X,Y),\theta^\star) \leq D(P(X,Y),\theta_S^\star) \leq \frac{D(P(X,Y),\theta^\star)}{P(\bar{U})}. \tag{11}$$

That is, the divergence minimised by supervised learning, $D(P(X,Y),\theta)$, evaluated at the parameters which minimise the semi-supervised divergence, $\theta_S^\star$, can be upper and lower bounded as a function of said divergence evaluated at its own optima, $\theta^\star$.

Proof The lower bound

$$D(P(X,Y),\theta^\star) \leq D(P(X,Y),\theta_S^\star)$$

is true by the definition of θ^* - it is the minimiser of $D(P(X,Y),\theta)$, and so any other value of θ must result in a greater than or equal divergence.

The upper bound can be derived as follows. Consider the term $D(P(X,Y),\theta_S^\star)P(\bar{U})$. Using Equation (9) evaluated at $\theta = \theta_S^\star$ we can see the following,

$$D(P(X,Y),\theta_S^\star)P(\bar{U}) \leq D(P(X,Z),\theta_S^\star).$$

Given the definition of θ_S^* we can further see that

$$D(P(X,Z),\theta_S^\star) \leq D(P(X,Z),\theta^\star).$$

And using Equation (10) evaluated at $\theta = \theta^\star$,

$$D(P(X,Z),\theta^\star) \leq D(P(X,Y),\theta^\star).$$

Hence, utilising all three of these inequalities in that order,

$$\begin{aligned} D(P(X,Y),\theta_S^\star)P(\bar{U}) &\leq D(P(X,Z),\theta_S^\star) \\ &\leq D(P(X,Z),\theta^\star) \\ &\leq D(P(X,Y),\theta^\star) \end{aligned}$$

we see that

$$D(P(X,Y),\theta_S^\star)P(\bar{U}) \leq D(P(X,Y),\theta^\star).$$

By dividing through by $P(\bar{U})$ we achieve our upper bound in Equation (11), that is,

$$D(P(X,Y),\theta_S^\star) \leq \frac{D(P(X,Y),\theta^\star)}{P(\bar{U})}.$$

■

Thus, we can place bounds on divergence $D(P(X,Y),\theta)$ evaluated at θ^*_S in terms of the proportion of $P(\bar{U})$, and $D(P(X,Y),\theta^*)$. One immediate observation is that if $D\big(P(X,Y,\theta^\star\big) = 0$, then $D\big(P(X,Y),\theta^\star_S\big) = 0$. Thus, if the true distribution lies within the family of distributions expressible by our model, then the optima intersect regardless of $P(U)$, as confirmed by Cozman et al. (2003). Conversely, if $D\big(P(X,Y),\theta^\star\big) > 0$ then our bounds loosen as $P(U)$ grows, and the rate of this depends on how well matched our model is to the data - if they are very similar then the bound grows slowly, whereas if they are different it may grow much faster. This confirms earlier results (see Section 2.2 in Zhu, 2005, for a summary), and builds on them by providing explicit bounds on how rapidly performance may degrade.

The overall conclusion is that performing ML semi-supervised learning in the manner of Equation (3) forces us to make a trade off. We can rarely evaluate KL divergences directly, and must use estimators whose variance is inversely proportional to N (MacKay, 2003). By including unlabelled data we can decrease this source of uncertainty. However in doing so we weaken our bounds, introducing a new source of error. This provides a complementary reinterpretation of the results noted by Cozman et al. (2003).

As our bounds weaken then, how does our solution degrade? We now show that the supervised divergence, evaluated at the ML semi-supervised optima, grows monotonically with the proportion of unlabelled samples.

Theorem 8 *Subject to the conditions in 3.1, let us define two distributions $P_1(X,Z)$ and $P_2(X,Z)$, and corresponding models $P_1(X,Z|\theta)$ and $P_2(X,Z|\theta)$. These distributions shall differ from one another only in terms of the probability that $Z = U$; that is, $P_1(X,Y) = P_2(X,Y)$ and $P_1(X,Y|\theta) = P_2(X,Y|\theta)$ (which in turn implies $P_1(X) = P_2(X)$ and $P_1(X|\theta) = P_2(X|\theta)$). We shall assume that distribution $P_2(X,Z)$ has a greater chance of an unlabelled sample, and so $P_2(U) > P_1(U)$.*

Define the optima $\theta^\star_{S1}$ and $\theta^\star_{S2}$ as

$$\theta^\star_{S1} = \arg\min_\theta D(P_1(X,Z),\theta), \quad \theta^\star_{S2} = \arg\min_\theta D(P_2(X,Z),\theta). \tag{12}$$

It follows that

$$D(P(X,Y),\theta^\star_{S1}) \leq D(P(X,Y),\theta^\star_{S2}) \tag{13}$$

and

$$D(P(Y|X),\theta^\star_{S1}) \leq D(P(Y|X),\theta^\star_{S2}). \tag{14}$$

Proof By definition,

$$D(P_1(X,Z),\theta^*_{S1}) \leq D(P_1(X,Z),\theta^*_{S2}).$$

We can expand both these divergences to rewrite this expression as follows;

$$\begin{aligned} &P_1(\bar{U})D(P(X,Y),\theta^\star_{S1}) + P_1(U)D(P(X),\theta^\star_{S1}) \\ &\leq P_1(\bar{U})D(P(X,Y),\theta^\star_{S2}) + P_1(U)D(P(X),\theta^\star_{S2}) \end{aligned}$$

Rearranging this expression to isolate $D(P(X), \theta^\star_{S1}) - D(P(X), \theta^\star_{S2})$ gives us

$$D(P(X), \theta^\star_{S1}) - D(P(X), \theta^\star_{S2}) \leq \frac{P_1(\bar{U})}{P_1(U)} \left(D(P(X,Y), \theta^\star_{S2}) - D(P(X,Y), \theta^\star_{S1}) \right). \tag{15}$$

We shall utilise this term later.

Now examining the divergences associated with the distribution $P_2(X, Z)$, by the definition given in Equation (12) we see that

$$D(P_2(X,Z), \theta^*_{S2}) \leq D(P_2(X,Z), \theta^*_{S1}).$$

This can be expanded as before,

$$\begin{aligned} &P_2(\bar{U}) D(P(X,Y), \theta^\star_{S2}) + P_2(U) D(P(X), \theta^\star_{S2}) \\ &\leq P_2(\bar{U}) D(P(X,Y), \theta^\star_{S1}) + P_2(U) D(P(X), \theta^\star_{S1}), \end{aligned}$$

and $D(P(X), \theta^\star_{S1}) - D(P(X), \theta^\star_{S2})$ once again isolated,

$$\frac{P_2(\bar{U})}{P_2(U)} \left(D(P(X,Y), \theta^\star_{S2}) - D(P(X,Y), \theta^\star_{S1}) \right) \leq D(P(X), \theta^\star_{S1}) - D(P(X), \theta^\star_{S2}). \tag{16}$$

Combining Equation (15) and Equation (16) to eliminate $D(P(X), \theta^\star_{S1}) - D(P(X), \theta^\star_{S2})$ gives us

$$\begin{aligned} &\frac{P_2(\bar{U})}{P_2(U)} \left(D(P(X,Y), \theta^\star_{S2}) - D(P(X,Y), \theta^\star_{S1}) \right) \\ &\leq \frac{P_1(\bar{U})}{P_1(U)} \left(D(P(X,Y), \theta^\star_{S2}) - D(P(X,Y), \theta^\star_{S1}) \right). \end{aligned}$$

Gathering together similar divergences, this implies that

$$\left(\frac{P_1(\bar{U})}{P_1(U)} - \frac{P_2(\bar{U})}{P_2(U)} \right) D(P(X,Y), \theta^\star_{S1}) \leq \left(\frac{P_1(\bar{U})}{P_1(U)} - \frac{P_2(\bar{U})}{P_2(U)} \right) D(P(X,Y), \theta^\star_{S2})$$

As we know that $P_2(U) > P_1(U)$, and so $P_2(\bar{U}) < P_1(\bar{U})$, it follows that $P_2(U)P_1(\bar{U}) > P_1(U)P_2(\bar{U})$, which in turn implies

$$\frac{P_1(\bar{U})}{P_1(U)} - \frac{P_2(\bar{U})}{P_2(U)} > 0.$$

As it is positive we may cancel this term out without altering the inequality, indicating that

$$D(P(X,Y), \theta^\star_{S1}) \leq D(P(X,Y), \theta^\star_{S2})$$

proving Equation (13).

To prove Equation (14), note that if we take Equation (15), multiply though by $P_1(U)$, and then use some simple algebra to gather all terms relating to the marginal divergence together, it is equivalent to stating

$$D(P(X), \theta^\star_{S1}) - D(P(X), \theta^\star_{S2}) \leq P_1(\bar{U}) \left(D(P(Y|X), \theta^\star_{S2}) - D(P(Y|X), \theta^\star_{S1}) \right).$$

Similarly, Equation (16) can be rearranged as

$$P_2(\bar{U})\left(D(P(Y|X),\theta^\star_{S2})-D(P(Y|X),\theta^\star_{S1})\right)\leq D(P(X),\theta^\star_{S1})-D(P(X),\theta^\star_{S2}).$$

Combining these two, we see that

$$\begin{aligned}&P_2(\bar{U})\left(D(P(Y|X),\theta^\star_{S2})-D(P(Y|X),\theta^\star_{S1})\right)\\&\leq P_1(\bar{U})\left(D(P(Y|X),\theta^\star_{S2})-D(P(Y|X),\theta^\star_{S1})\right).\end{aligned}$$

Gathering together terms, this rearranges to

$$\left(P_1(\bar{U})-P_2(\bar{U})\right)D(P(Y|X),\theta^\star_{S1})\leq\left(P_1(\bar{U})-P_2(\bar{U})\right)D(P(Y|X),\theta^\star_{S2})$$

which, given $P_2(U)>P_1(U)$, and hence $P_2(\bar{U})<P_1(\bar{U})$, implies

$$D(P(Y|X),\theta^\star_{S2})\geq D(P(Y|X),\theta^\star_{S1})$$

proving Equation (14). ■

This observation seems intuitively reasonable. As $P(U)$ grows the model is increasingly penalised by large values of $D(P(X),\theta)$, and so seeks to minimise this at the expense of letting $D(P(Y|X),\theta)$ get larger. However, if our end goal is to create a classifier then this result may give us cause to reconsider - adding unlabelled data not only weakens our bounds on the joint divergence, but asymptotically can only worsen (or at best leave unchanged) the conditional divergence.

Thus, we can now conclude several things about the asymptotic optimum of the ML semi-supervised learning problem. Firstly, due to the observation of monotonicity, the divergence $D(P(X,Y),\theta^*_S)$ is upper bounded by $D(P(X,Y),\theta^*_U)$, confirming Yang and Priebe (2011). Secondly, that if we were to increase the quantity of unlabelled data, it will tend towards this approaching equality as $P(U)$ tends towards 1. Finally, it will do so monotonically - raising the proportion of unlabelled data will never decrease $D(P(X,Y),\theta^*_S)$ or $D(P(Y|X),\theta^*_S)$.

This result initially seems to contradict that of Cozman and Cohen (2006), where they gave an example of a ML semi-supervised learning process where despite the model not fitting the underlying distribution, adding unlabelled data asymptotically improved the decision boundary. We point out though that their measure of how well the boundary fits is based on the error rate, not the KL divergence. while it is true that minimising the conditional KL divergence will typically reduce the error rate this is not an absolute rule (and indeed forms a set of bounds). We would postulate that this is an example of a case where the divergence rises but the classification rate improves.

Finally, it makes sense to more closely examine the final solution arrived at as $P(U)\to 1$, in a similar manner to that discussed by Yang and Priebe (2011). In particular, we wish to confirm that their result extend beyond identifiable models, and shall show that where there is a choice between multiple sets of parameters which minimise the unsupervised divergence $D(P(X),\theta)$, the semi-supervised divergence minimised by ML learning is upper bounded by the one set of these parameters which best minimises $D(P(Y|X),\theta)$. This proof is largely similar to that showing monotonicity but is included here for completeness.

Theorem 9 *Subject to the conditions in 3.1, define the optimum unsupervised parameters $\theta_U^\star$ to be any parameters which meet these requirements:*

$$\theta_U^\star = \arg\min_\theta D(P(Y|X),\theta) \quad \textit{subject to} \quad D(P(X),\theta_U^\star) = \min_{\theta'} D(P(X),\theta')$$

It can be shown that provided $P(\bar{U}) \neq 0$,

$$D(P(X,Y),\theta_S^\star) \leq D(P(X,Y),\theta_U^\star) \tag{17}$$

and

$$D(P(Y|X),\theta_S^\star) \leq D(P(Y|X),\theta_U^\star) \tag{18}$$

Proof By definition,

$$D(P(X,Z),\theta_S^\star) \leq D(P(X,Z),\theta_U^\star). \tag{19}$$

The standard semi-supervised divergence can be expanded as follows,

$$D(P(X,Z)|\theta) = P(\bar{U})D\big(P(X,Y),\theta\big) + P(U)D\big(P(X),\theta\big).$$

As such, we can rewrite Equation (19) as follows

$$\begin{aligned} &P(\bar{U})D(P(X,Y),\theta_S^\star) + P(U)D(P(X),\theta_S^\star) \\ &\leq P(\bar{U})D(P(X,Y),\theta_U^\star) + P(U)D(P(X),\theta_U^\star). \end{aligned}$$

If we subtract $P(U)D(P(X),\theta_U^\star)$ from both sides this becomes

$$\begin{aligned} &P(\bar{U})D(P(X,Y),\theta_S^\star) + P(U)\left(D(P(X),\theta_S^\star) - D(P(X),\theta_U^\star)\right) \\ &\leq P(\bar{U})D(P(X,Y),\theta_U^\star). \end{aligned}$$

However, by definition, $D(P(X),\theta_S^\star) \geq D(P(X),\theta_U^\star)$, implying

$$P(\bar{U})D(P(X,Y),\theta_S^\star) \leq P(\bar{U})D(P(X,Y),\theta_U^\star)$$

and hence Equation (17) directly follows by dividing through by $P(\bar{U})$. Equation (18) follows similarly by noting that Equation (17) implies

$$D(P(Y|X),\theta_S^\star) + D(P(X),\theta_S^\star) \leq D(P(Y|X),\theta_U^\star) + D(P(X),\theta_U^\star).$$

If we subtract $D(P(X),\theta_U^\star)$ from both sides we find that

$$D(P(Y|X),\theta_S^\star) + D(P(X),\theta_S^\star) - D(P(X),\theta_U^\star) \leq D(P(Y|X),\theta_U^\star)$$

and again note that by definition, $D(P(X),\theta_S^\star) \geq D(P(X),\theta_U^\star)$, and hence

$$D(P(Y|X),\theta_S^\star) \leq D(P(Y|X),\theta_U^\star)$$

directly follows, proving Equation (18).

■

Note that the above derivation could have proceeded in exactly the same manner with $\theta^\star_U$ chosen to be any parameters for which $D(P(X), \theta^\star_U) = \min_{\theta'} D(P(X), \theta')$. However, by choosing $\theta^\star_U$ to be the parameters which also minimised $D(P(Y|X), \theta^\star_U)$ we attain as tight a bound as possible.

For many models there will be only one set of parameters which minimise $D(P(X), \theta)$, and so this is not an issue. However for others this will not be the case. For example, many mixture models contain mixture components which are identical save for the class they are assigned to. In these cases, specifying that $\theta^\star_U$ have the lowest conditional divergence amongst those set of parameters which have the minimum marginal divergence allows us to choose the best combination of class assignment for each mixture component given their other parameters, strengthening slightly the conclusions of Yang and Priebe (2011).

We can now rewrite our global bounds as follows:

$$D(P(X,Y), \theta^\star) \leq D(P(X,Y), \theta^\star_S) \leq \min\left(\frac{D(P(X,Y), \theta^\star)}{P(\bar{U})}, D(P(X,Y), \theta^\star_U)\right).$$

Assuming that $D(P(Y|X), \theta^\star_U) < \infty$, which will be the case provided $P(Y|X, \theta^\star_U)$ does not assign zero probability to any Y given any X, this gives tighter performance bounds as $P(\bar{U}) \to 0$.

3.5 Summary

This ends our theoretical examination of performing ML learning on a partially labelled data set. Overall we can conclude the following;

- When we introduce unlabelled data into our likelihood expression, we change the divergence being minimised, from $D(P(X,Y), \theta)$ to $D(P(X,Z), \theta)$.

- We can form a set of upper and lower bounds on $D(P(X,Y), \theta)$ using $D(P(X,Z), \theta)$ for a given θ, namely

  $$D(P(X,Z), \theta) \leq D(P(X,Y), \theta) \leq \frac{D(P(X,Z), \theta)}{P(\bar{U})}.$$

  The lower bound becomes tight if $D(P(Y|X), \theta)$ is equal to 0, that is, if our model is fits to the conditional distribution well.

- If we find the parameters $\theta^\star_S$ which minimise the standard semi-supervised divergence $D(P(X,Z), \theta)$, then these are linked to the parameters $\theta^\star$ which minimise $D(P(X,Y), \theta)$ using the expression

  $$D(P(X,Y), \theta^\star) \leq D(P(X,Y), \theta^\star_S) \leq \frac{D(P(X,Y), \theta^\star)}{P(\bar{U})},$$

  that is, our supervised divergence evaluated at the standard semi-supervised minima may exceed the supervised minima by a factor of $1/(P(\bar{U}))$. Where there is a large quantity of unlabelled data this factor may be very high.

- $D(P(X,Y), \theta^\star_S)$ grows monotonically with the proportion of unlabelled data $P(U)$. Moreover, it is the term $D(P(Y|X), \theta^\star_S)$ which grows, indicating that we can expect classification results to remain steady or worsen.

Taken together, this gives a clear indication of the problem we face conducting generative semi-supervised ML learning, and gives novel bounds on the asymptotic performance achievable.

4. Unbiased Generative Semi-Supervised Learning

Having investigated the properties of $D(P(X,Z),\theta)$, it is clear that if we wish to minimise $D(P(X,Y),\theta)$, it is better we find an unbiased likelihood estimator. From examination of the form of the supervised divergence, we propose the following.

Theorem 10 *Subject to the conditions in 3.1, and provided $P(\bar{U}) > 0$, the expression*

$$\arg\max_\theta \prod_{i=1}^{N_L} P(y_i|x_i,\theta)\Big(\prod_{i=1}^{N} P(x_i|\theta)\Big)^{N_L/N} \tag{20}$$

returns a set of parameters which minimise an asymptotically unbiased estimator of the divergence $D(P(X,Y),\theta)$.

Proof The divergence $D(P(X,Y),\theta)$ is exactly equivalent to the following:

$$D(P(Y|X),\theta) + D(P(X),\theta). \tag{21}$$

We draw samples $(x_i, y_i)_{i=1,\ldots,N_L}$ and $(x_i)_{i=N_L+1,\ldots,N}$. Assuming that as $N \to \infty$, $N_L/N \to P(\bar{U})$, we can use these to construct an asymptotically unbiased estimator of the divergence Equation (21),

$$\left(\frac{1}{N_L}\sum_{i=1}^{N_L}\log\left(\frac{P(y_i|x_i)}{P(y_i|x_i,\theta)}\right) + \frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{P(x_i)}{P(x_i|\theta)}\right)\right). \tag{22}$$

Disregarding all terms which are not a function of θ gives the expression

$$\left(\frac{-1}{N_L}\sum_{i=1}^{N_L}\log\left(P(y_i|x_i,\theta)\right) + \frac{-1}{N}\sum_{i=1}^{N}\log\left(P(x_i|\theta)\right)\right). \tag{23}$$

Multiplying this by $-N_L$ and taking the antilog yields

$$\prod_{i=1}^{N_L} P(y_i|x_i,\theta)\prod_{i=1}^{N} P(x_i|\theta)^{N_L/N}.$$

This is the quantity which is maximised in Equation (20). As the log function is monotonic, the parameters which maximise this will minimise Equation (23) (due to the multiplication by $-N_L$). Minimising Equation (23) is equivalent to minimising Equation (22) as they only differ by terms that are constant with respect to the parameters. And Equation (22) is an unbiased estimator of Equation (21). Hence, the parameters returned by Equation (20) are equivalent to those which minimise an unbiased estimator of $D(P(X,Y),\theta)$. ■

A special case occurs when $P(\bar{U}) = 0$. This corresponds to Equation (22) where N_L is fixed while $N_U \to \infty$,

$$\arg\min_{\theta} \frac{1}{N_L} \sum_{i=1}^{N_L} \log\left(\frac{P(y_i|x_i)}{P(y_i|x_i,\theta)}\right) + D(P(X),\theta) \tag{24}$$

which estimates the marginal component of the divergence exactly while using the available labelled data to estimate the conditional component as best possible.

Our term Equation (20) is somewhat similar to the form of Equation 3 presented by McCallum et al. (2006), which was further investigated by Druck et al. (2007), but with the exponents of the conditional and generative components of the equation set by the ratio of labelled to unlabelled data, rather than being found by cross validation. Moreover, our purpose in using an equation of this form is different; we wish to fit a generative model, not a classifier. Rosset et al. (2005) has also previously noted that performance can be improved by requiring certain expectations in the labelled and unlabelled data set match. However, they enforced this as a strict requirement, rather than using it to find an unbiased likelihood estimate as we do. Nigam et al. (2000) implement down-weighting of the log likelihood of all unlabelled elements, by a factor which is set using cross validation. As the marginal likelihood of the labelled samples is not re-weighed this also produces a biased estimator of the joint likelihood.

An argument might be made that a biased estimator which is tuned using cross validation has the potential to outperform the proposed unbiased objective function. While there is certainly merit to this, we would respond that the parameter tuning inherent in cross validation can increase the amount of time spent training dramatically, and that it requires a large enough corpus of labelled data that a holdout set can be safely put aside to validate with. Our objective function provides a simple, principled alternative, applicable in cases where such restrictions prevent cross validation, as well as others.

4.1 Estimator Variance

Note that we can already generate an asymptotically unbiased estimate by using the labelled data alone. Unlabelled samples are only of value if they make this process more reliable, so it is worth investigating the uncertainty of this estimator. Consider the variance V of Equation (22), where the variance is taken w.r.t. the probability of the possible data sets we may have observed

$$V = \mathrm{Var}\left(\frac{1}{N_L}\sum_{i=1}^{N_L}\log\left(\frac{P(y_i|x_i)}{P(y_i|x_i,\theta)}\right) + \frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{P(x_i)}{P(x_i|\theta)}\right)\right). \tag{25}$$

We can expand Equation (25) as

$$\begin{aligned} V &= \mathrm{Var}\left(\frac{1}{N_L}\sum_{i=1}^{N_L} L_{y|x}(i)\right) + \mathrm{Var}\left(\frac{1}{N}\sum_{i=1}^{N} L_x(i)\right) \\ &\quad +2\,\mathrm{Cov}\left(\frac{1}{N_L}\sum_{i=1}^{N_L} L_{y|x}(i), \frac{1}{N}\sum_{i=1}^{N} L_x(i)\right) \end{aligned} \tag{26}$$

where for ease of notation we have defined

$$L_{y|x}(i) \equiv \log\left(\frac{P(y_i|x_i)}{P(y_i|x_i,\theta)}\right), \quad L_x(i) \equiv \log\left(\frac{P(x_i)}{P(x_i|\theta)}\right).$$

Using standard identities for variances and covariances,[2] and taking advantage of our samples being iid, we can expand Equation (26) as follows

$$V = \frac{1}{N_L}\operatorname{Var}\left(L_{Y|X}\right) + \frac{1}{N}\operatorname{Var}\left(L_X\right) + \frac{2}{N}\operatorname{Cov}\left(L_{Y|X}, L_X\right) \tag{27}$$

where we have now defined

$$L_{Y|X} \equiv \log\left(\frac{P(Y|X)}{P(Y|X,\theta)}\right), \quad L_X \equiv \log\left(\frac{P(X)}{P(X|\theta)}\right).$$

As such Equation (27) gives us the variance of Equation (22) in terms of a relationship between the distributions $P(X,Y)$, $P(X,Y|\theta)$, N_L and N_U. Clearly $V \to 0$ as $N_L \to \infty$. However we are also interested in the case where we increase N_U while holding N_L steady, corresponding to $P(\bar{U}) = 0$. By inspection, as $N_U \to \infty$, $V \to \frac{1}{N_L}\operatorname{Var}\left(L_{Y|X}\right)$ (as one might expect from examination of Equation (24)), so the question becomes whether this reduces V. Remembering that $N = N_L + N_U$, the derivative[3] of Equation (27) with respect to N (and hence N_U since N_L is fixed) is

$$\frac{dV}{dN} = \frac{-1}{N^2}\operatorname{Var}\left(L_X\right) - \frac{2}{N^2}\operatorname{Cov}\left(L_{Y|X}, L_X\right).$$

[2] The simplification of the variance terms is intuitively obvious and a standard result - for any random variable, we expect the variance of the arithmetic mean of a set of observations to be the variance of the variable itself, divided by the number of observations made (see MacKay, 2003). However the covariance term is perhaps a little more surprising, as it has no dependence on N_L. This is due to the iid nature of the data, which implies a covariance of zero between different samples. As such we can derive the following:

$$\begin{aligned}
&\operatorname{Cov}\left(\frac{1}{N_L}\sum_{i=1}^{N_L} L_{y|x}(i), \frac{1}{N}\sum_{i=1}^{N} L_x(i)\right)\\
&= \frac{1}{NN_L}\sum_{i=1}^{N_L}\sum_{j=1}^{N}\operatorname{Cov}\left(L_{y|x}(i), L_x(j)\right)\\
&= \frac{1}{NN_L}\sum_{i=1}^{N_L}\sum_{\substack{j=1\\ j\neq i}}^{N}\operatorname{Cov}\left(L_{y|x}(i), L_x(j)\right) + \frac{1}{NN_L}\sum_{i=1}^{N_L}\operatorname{Cov}\left(L_{y|x}(i), L_x(i)\right)\\
&= 0 + \frac{1}{NN_L}\sum_{i=1}^{N_L}\operatorname{Cov}\left(L_{y|x}(i), L_x(i)\right)\\
&= \frac{1}{NN_L}\sum_{i=1}^{N_L}\operatorname{Cov}\left(L_{Y|X}, L_X\right)\\
&= \frac{1}{N}\operatorname{Cov}\left(L_{Y|X}, L_X\right)
\end{aligned}$$

which eliminates N_L and so gives us the stated form of the covariance.

[3] Strictly N_U is discrete, and does not have a derivative. However the expression is monotonic in N_U, and so the overall trend will be the same if this constraint is relaxed.

For Equation (25) to decrease as N_U increases this quantity must be negative. This is the case iff

$$\mathrm{Cov}\left(L_{Y|X}, L_X\right) \geq \frac{-\mathrm{Var}\left(L_X\right)}{2}.$$

The conclusion is that even if N_L is fixed our method of including unlabelled data reduces the variance of our estimator provided $\mathrm{Cov}\left(L_{Y|X}, L_X\right)$ is above a lower bound, proportional to $\mathrm{Var}\left(L_X\right)$. Perhaps surprisingly this bound is negative, indicating they may be slightly anti-correlated. We feel this is a sufficiently weak criteria for our scheme to find application across a variety of data sets.

5. Empirical Demonstration

We now examine the performance of the objective function given in Section 4 on real world data sets, compared to the standard semi-supervised learning, supervised learning, and several other alternative semi-supervised techniques. To maximally highlight the effect of mismatch between the model and true distribution, a simple marginal distribution consisting of a single axis aligned Gaussian was chosen to model each class.

The following six learning schemes were tested with this model: our unbiased semi-supervised expression (**SSunb**), that is, the natural log of Equation (20); the log likelihood of the labelled data (**LL**), that is, Equation (1); the log likelihood of the standard (biased) semi-supervised expression (**SSb**), that is, the natural log of Equation (3); the log likelihood of the standard semi-supervised expression plus an Entropy Regularisation term (Grandvalet and Bengio, 2006) with the parameter λ set by 5 fold cross validation, selecting the λ with the lowest holdout set error rate (**ERer**); Entropy Regularisation as before, except cross validation is carried out on the log likelihood of the holdout set (**ERnll**); the semi-supervised equivalent of Multi Conditional learning (as investigated in Druck et al., 2007), again cross validating hyper parameters once on error rate (**MCer**) and once on log likelihood (**MCnll**); and the log likelihood of the standard semi-supervised expression plus an Expectation Regularisation (Mann and McCallum, 2007) term (**XR**), with the trade off parameter set (after some experimentation) as in the original paper to the equivalent of 10 times the number of labelled samples; Additionally, for the position parameter μ of each Gaussian a penalty term $-C||\mu||^2$ was added onto each objective function with C set to a small constant ($\approx 10^{-5}$).

We would point out that many of these learning schemes were originally designed for use with a discriminative model. Here we are using them in a different manner, to augment the objective function during the learning of a generative model. They have been selected due to their reported good performance in improving discriminative learning, in the hope that this will counteract the bias introduced by the missing class information in the likelihood of the unlabelled samples.

We chose 7 data sets from the UCI repository (Frank and Asuncion, 2010); **Diabetes**, **Wine**, glass identification (**Glass**), blood transfusion (**Blood**) (Yeh et al., 2009), **Ecoli**, Haberman survival (**Haber**), and Pima Indian diabetes (**Pima**); and 2 from libsvm: SVM guide 1 (**SVMg**) (Hsu et al., 2003) and fourclass (**Four**) (Ho and Kleinberg, 1996). Due to computational constraints, data sets with > 3 classes had one or more merged to create 3 approximately equally sized groupings. Each axis of the data was transformed to lie in

data	SSunb	LL	SSb	MCer	MCnll	ERer	ERnll	XR
Diabetes	**3.36**	4.12	3.90	144	3.58	3.90	3.97	3.74
SVMg	0.379	0.417	1.18	99.4	**0.376**	1.23	1.19	1.16
Wine	19.4	58.4	23.0	67.2	21.4	24.7	24.7	**12.5**
Glass	23.7	40.9	23.5	213	26.3	22.7	**21.5**	21.5
Blood	**1.78**	2.27	3.01	77.2	2.08	3.06	3.06	2.65
ecoli	**8.80**	13.7	10.0	68.6	10.3	9.97	10.0	9.63
Haber	4.75	7.30	5.02	79.8	5.10	4.98	4.90	**4.59**
Pima	**3.60**	4.30	4.24	136	3.80	4.26	4.25	3.87
Four	**2.17**	2.22	2.25	37.3	2.19	2.33	2.32	2.23

Table 1: Overall mean negative log likelihood - best result for each data set shown in **bold**, second best underlined

the range $[-1, 1]$. Samples with missing attributes were excluded. Where a data set had a dedicated test set, this was used; otherwise, one fifth of the data was randomly separated a priori for this purpose.

A range of values of N_L and N_U were trialled. As a proportion of the total available training data, N_L varied from $[0.025, 0.05, 0.1, 0.2]$, and N_U from $[0.025, 0.05, 0.1, 0.2, 0.4, 0.8]$, with N_U being formed by discarding labels prior to training (for example, a test where $N_L = 0.05$ and $N_U = 0.4$ would indicate 0.45 of the available data was used for training, of which one ninth was labelled). For each repetition a random set of parameters was generated and used as the starting point for each of the above learning schemes. Each model was optimised by repeatedly alternating between a small number of iterations of downhill simplex search (Lagarias et al., 1998), followed by a large numbers of iterations of BFGS search (Nocedal and Wright, 1999), until convergence. This process was repeated 100 times for each combination of N_L and N_U values. The error rate and negative log likelihood of the test set was found for each solution. A selection of these results are shown here. Full results over all test sets are included in the appendix.

Note that we have purposefully used the same optimisation scheme for all objective functions - including **LL**, which has a closed form solution, and **SSb**, which can be optimised using expectation maximisation. Also note that for each repetition a single set of starting parameters was randomly generated, and then used to initialise every learning scheme investigated. The intent of this was to ensure that all variability encountered was solely due to the choice of objective function.

Table 1 shows the mean negative log likelihood for each data set, that is, the negative log likelihood averaged over all all repetitions of all values of N_L and N_U. For each data set, the minimum negative log likelihood is shown in bold, and the second smallest underlined. Our method **SSunb** achieved the lowest mean for 5 of the 9 data sets, and the second lowest for a further 3. **XR** proved the best for two data sets, and **MCnll** and **ERnll** for one each.

Mean error rates are shown in in Table 2. Our method performed best for 3 data sets and second best in a further 4, roughly equivalent to **LL** or **MCer** (without the need for the

data	SSunb	LL	SSb	MCer	MCnll	ERer	ERnll	XR
Diabetes	**0.283**	0.284	0.332	0.299	0.294	0.337	0.346	0.312
SVMg	0.0597	0.0597	0.189	**0.0572**	0.0678	0.193	0.175	0.171
Wine	**0.144**	0.163	0.190	0.238	0.162	0.230	0.259	0.218
Glass	0.483	**0.459**	0.539	0.470	0.501	0.545	0.555	0.525
Blood	0.277	0.277	0.367	**0.273**	0.292	0.372	0.369	0.309
ecoli	0.121	**0.104**	0.266	0.140	0.147	0.276	0.280	0.184
Haber	**0.298**	0.298	0.371	0.337	0.319	0.379	0.369	0.301
Pima	0.296	**0.292**	0.348	0.302	0.306	0.355	0.358	0.327
Four	0.249	0.250	0.260	**0.245**	0.250	0.281	0.274	0.259

Table 2: Overall mean error rate - best result for each data set shown in **bold**, second best underlined

latter's expensive cross validation). We point out that we are training a simple generative model, and so error rates reported are not directly comparable to previous work using more powerful / conditional models.

Figure 1 consists of four plots, showing how the mean negative log likelihood of the **Blood** data set varies as N_U is increased, for all four values of N_L tested. Error bars indicate a single standard deviation. Note how for small values of N_U all methods perform similarly, with some benefit from using unlabelled data. As N_U increases and the upper bound weakens, the methods begin to diverge - the **ER** methods, along with **SSb** and **XR** worsen consistently. **LL** remains approximately constant (as expected) and slightly larger than **SSunb**. **MCnll** sits somewhere between **LL** and **SSunb**, worsening a little as the proportion of unlabelled data grows. This qualitative description of the observed behaviour applies to a significant proportion of the results. The main exceptions to this trend were in the **Glass**, **Wine** and **Haber** data sets for small values of N_L, where competing methods (noticeably **XR**) performed better though this advantage tended to tail off as N_L grew - for example see Figure 3, which shows the Haberman data set performing very well under **XR** training.

Figure 2 shows how the mean error rate of the same data set varies. For small quantities of labelled data SSunb tends to tie with **LL**. However as the quantity of labelled data grows competing methods begin to out perform it. In general we found that the proposed unbiased method was not always best (most commonly being out performed by **XR**), but often very competitive. It also rarely showed degradation in behaviour as the quantity of unlabelled data was increased - as we would expect, given the manner in which it automatically downgrades the influence of additional unlabelled samples.

As well as looking at the mean log likelihood and error rates though, we believe another informative measure of the success of a semi-supervised algorithm is the raw frequency with which it out performs alternate methods. This gives an estimate of the probability that, should you include unlabelled data in your training data set, the performance of the algorithm will improve.

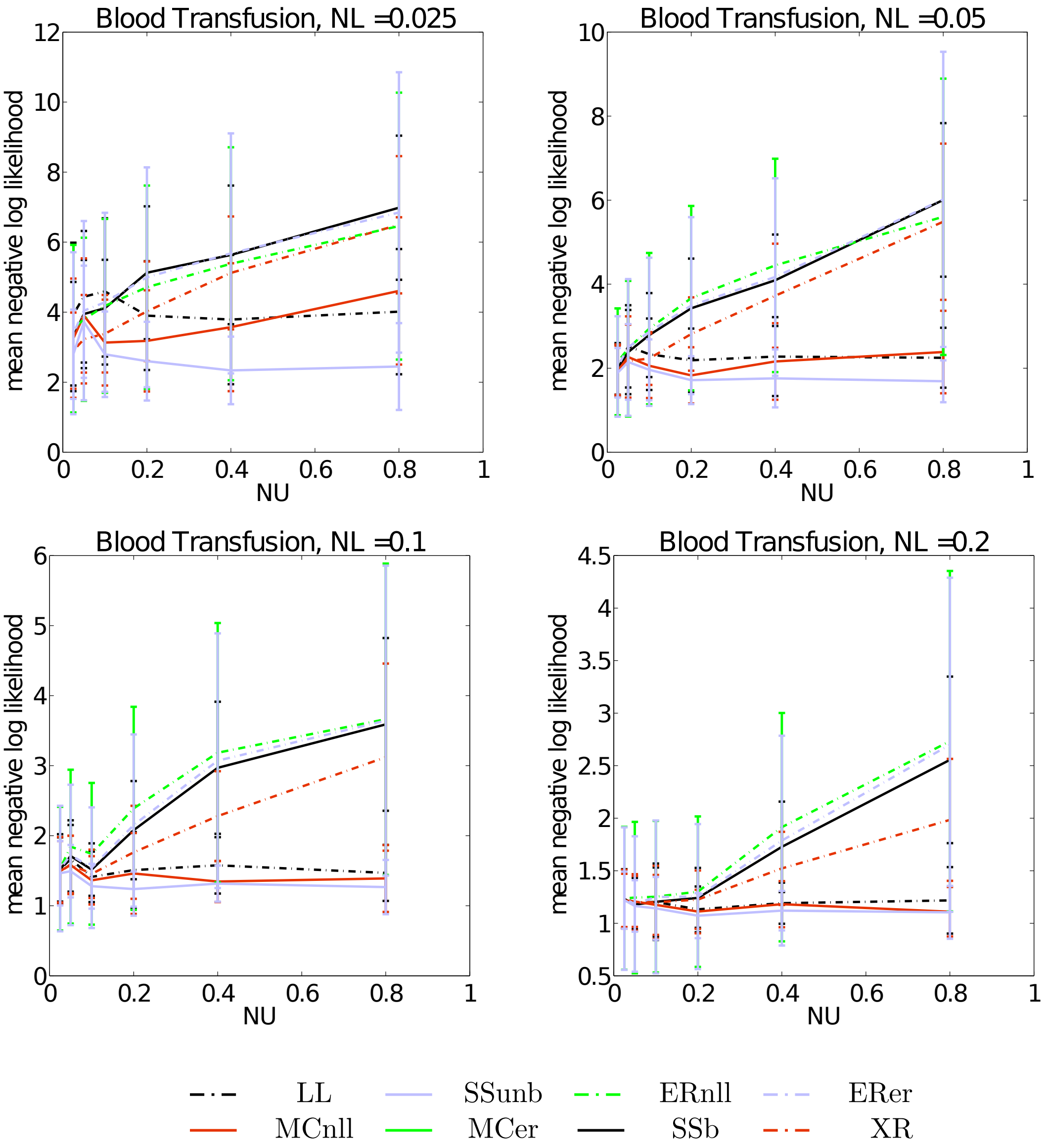

Figure 1: Four sample plots of the mean negative log likelihood of the **Blood** data set for a variety of values of N_L, as N_U grows. Note that $MCer$ is excluded, as it significantly underperformed and caused unfavourable axis scaling.

An example of this is shown in Figure 4, which shows the proportion of occasions each semi-supervised algorithm out performs supervised learning, where performance is measured in terms of the negative log likelihood of the test set. For this particular case our proposed unbiased estimator is consistently the superior one - on only one occasion does another algorithm (**MCnll**) outperform supervised learning with greater frequency. In general it was found that only when N_U is small that we typically saw other methods performing

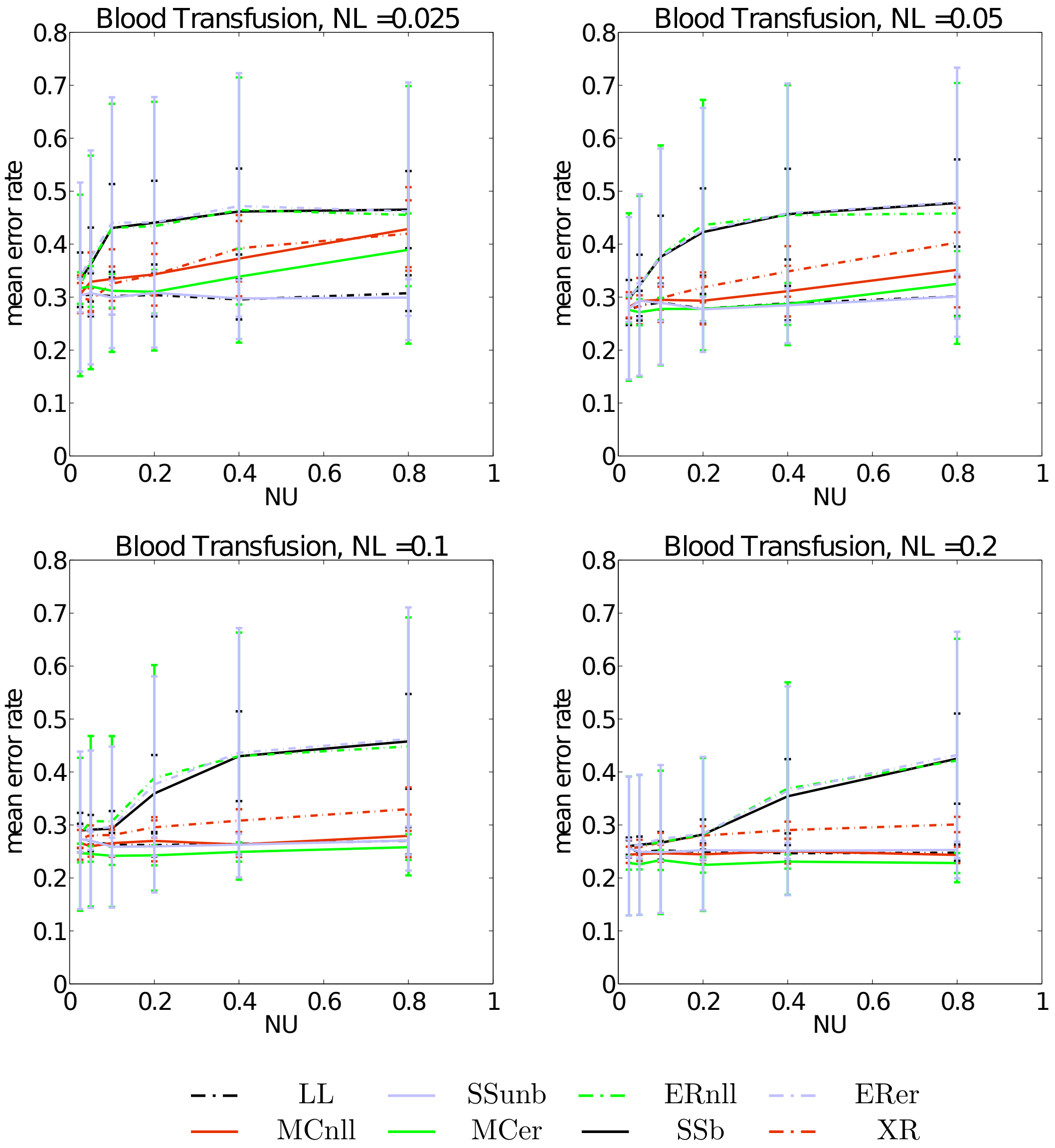

Figure 2: Four sample plots of the mean error rate of the **Blood** data set for a variety of values of N_L, as N_U grows.

better. What is also notable is how, while several other methods initially provide a bonus when N_U is small (where the proportion rises above 0.5, indicating that they were more likely than not to improve learning), they tend to degrade quite rapidly as unlabelled data is added, often making it more probable that they will worsen performance by the time $N_U = 0.8$. It was much rarer that our algorithm did this (one example occurring in the **Ecoli** data set with $N_L = 0.2$).

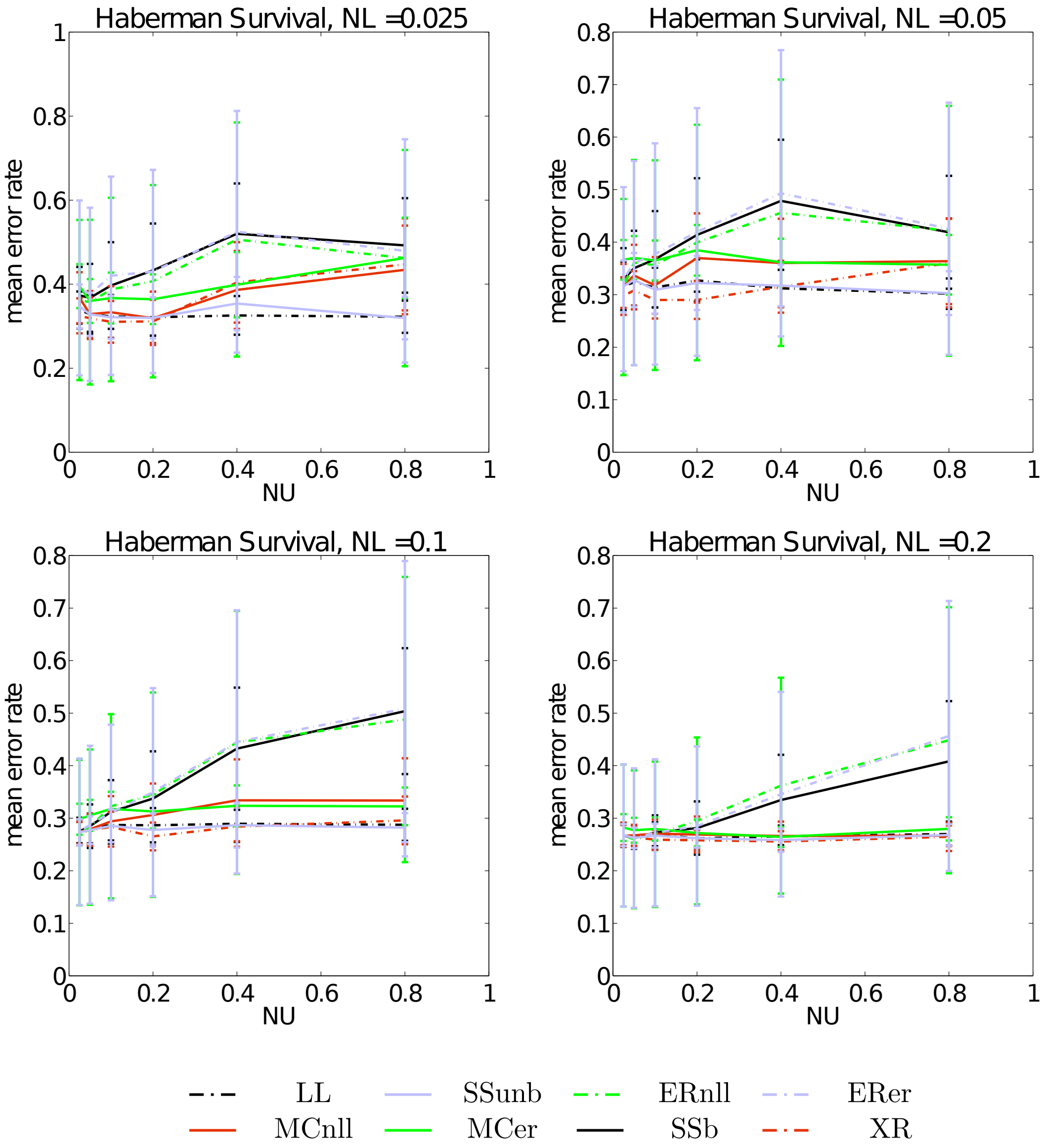

Figure 3: Four sample plots of the proportion of tests in which each semi-supervised learning scheme outperformed learning on the labelled data alone as measured by the negative log likelihood on the **Haberman** data set for a variety of values of N_L, as N_U grows.

Finally, Figure 5 shows the proportion of repetitions for which each semi-supervised algorithm reduced the error compared to supervised learning alone. Here our algorithm behaves much less impressively. With N_L set to its lowest value 0.025 it tends to be the better of the algorithms as N_U grows, but the proportion of occasions it provides a benefit is barely above 0.5. As the number of labelled data samples grows the two multi conditional

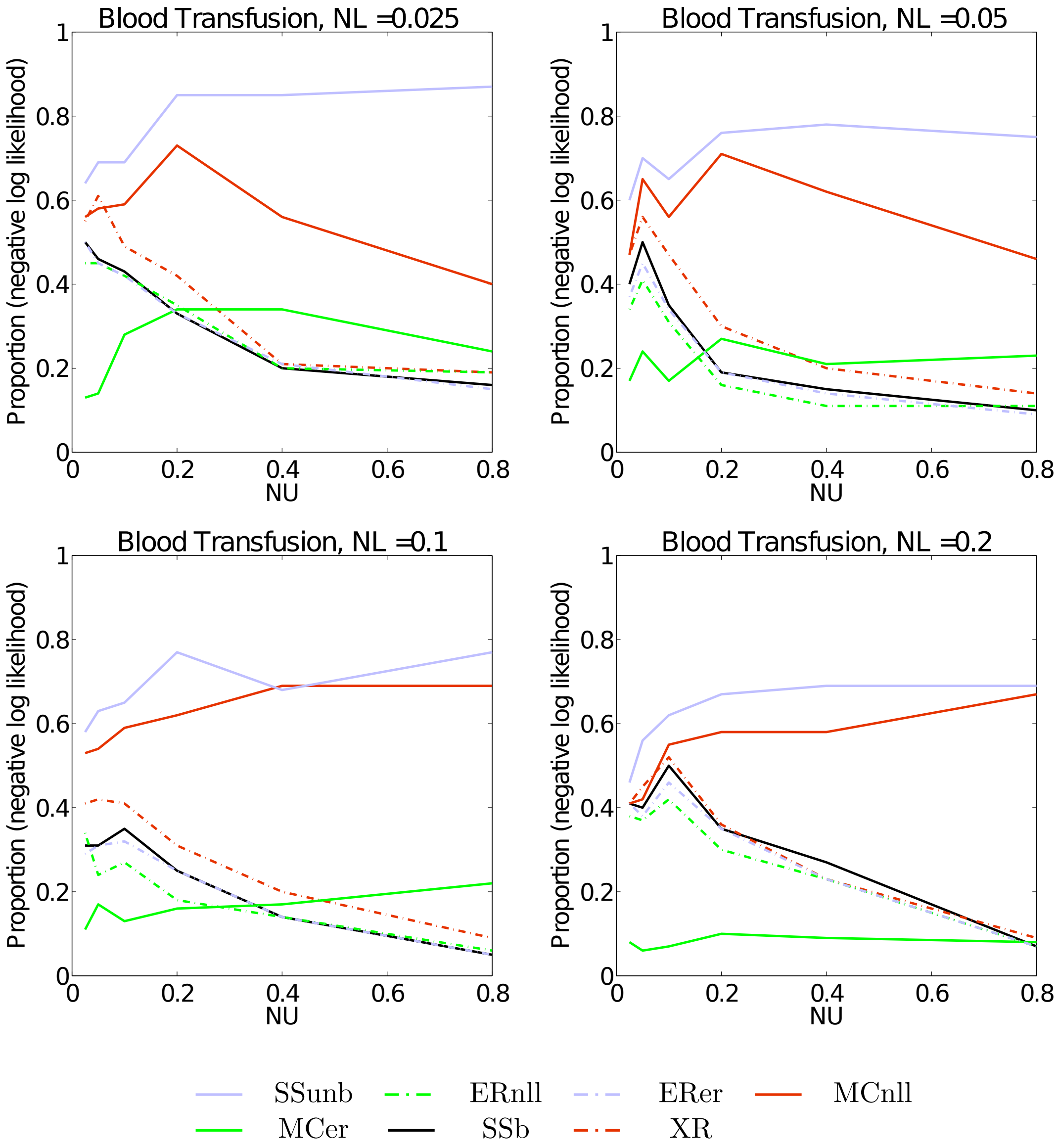

Figure 4: Four sample plots of the proportion of tests in which each semi-supervised learning scheme outperformed learning on the labelled data alone as measured by the negative log likelihood on the **Blood** data set for a variety of values of N_L, as N_U grows.

learning algorithms begin to out perform all others, especially when cross validated to reduce error.

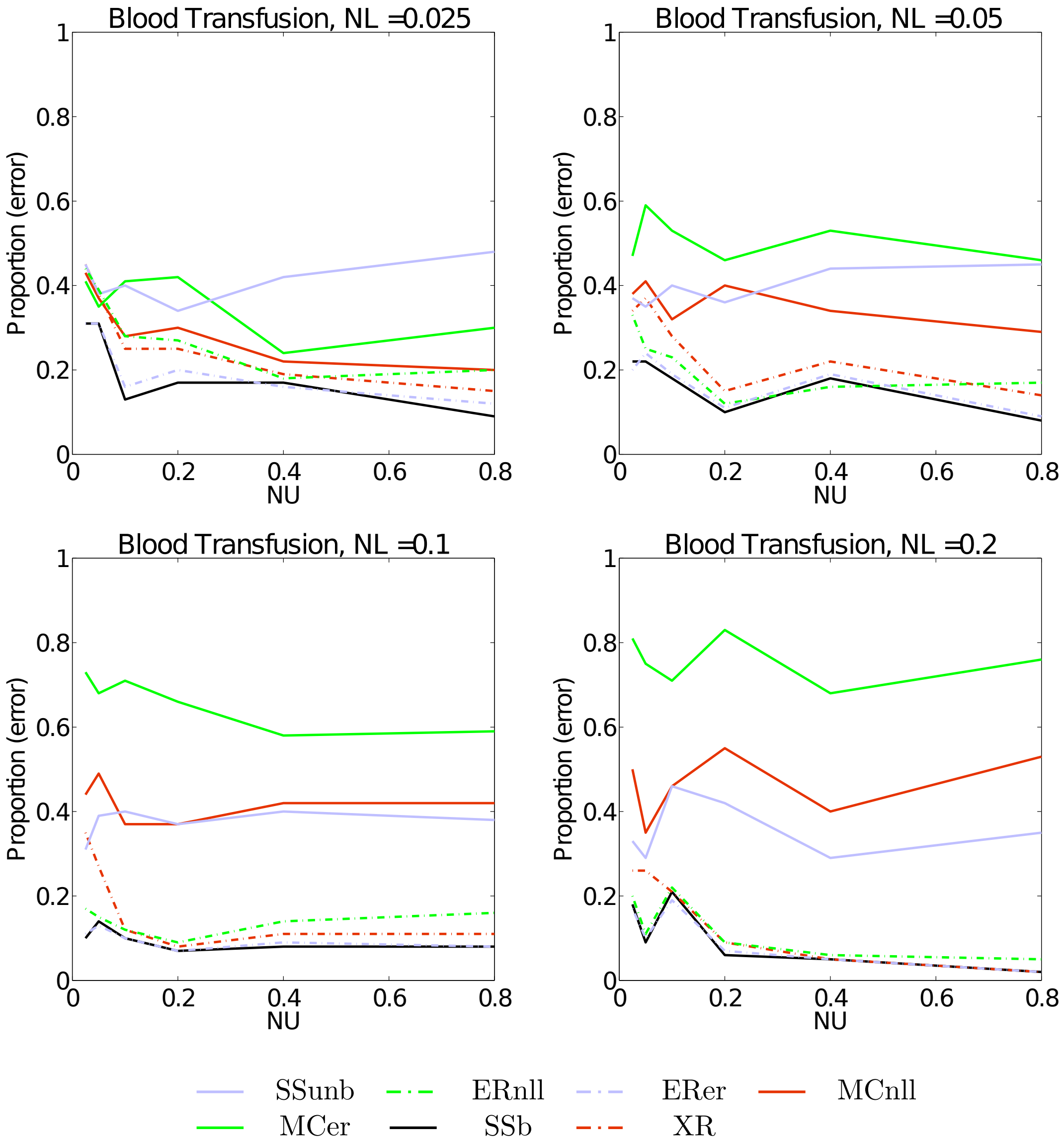

Figure 5: Four sample plots of the proportion of tests in which each semi-supervised learning scheme outperformed learning on the labelled data alone as measured by the error rate on the **Blood** data set for a variety of values of N_L, as N_U grows.

6. Conclusions

We have presented tighter bounds on the bias introduced when performing semi-supervised (as opposed to full supervised) maximum likelihood learning with a generative model using the standard technique. We also provide a new interpretation which gives an intuitive explanation for why the results are often poor with large amounts of unlabelled data.

Additionally, we have demonstrated a simple example of a new unbiased objective function which approximately minimises $KL\left(P(X,Y)||P(X,Y|\theta)\right)$. This method is no more computationally complex than simply augmenting the likelihood, demonstrates very good behaviour with even very large quantities of unlabelled data, and requires quite weak conditions on the correlation between the conditional and generative components of the likelihood to reduce the variance of our estimator.

Although not covered, much of the analysis presented here likely to be applicable to regression problems as well as classification ones. We leave this as an avenue for possible future work.

Acknowledgments

This work was supported by the EPSRC.

Appendix A. Supplementary Results Of Unbiased Semi-Supervised Training

There are two measures of performance we examine, evaluated over a holdout set:

- The error rate
- The negative log likelihood

For each of these measures two statistics are calculated:

- The mean performance over all repetitions, and the associated variance.
- The proportion of repetitions in which the performance was better than that achieved using the labelled data alone.

The former gives an approximate measure of 'average risk', according to whether we consider risk in terms of misclassification rate (for example when designing a classification algorithm) or negative log likelihood (such as when building a compression algorithm, say). The latter tells us, for each measure of 'risk', whether or not including unlabelled data has improved or worsened our performance.

Multi conditional learning, when cross validated according to error rate, often gave extremely bad negative log likelihood results. This caused unfavourable scaling of the axis, making other results indistinguishable. As such, the mean negative log likelihood results of **MCer** have been separated out and plotted alone.

See the main body of the text for such as abbreviations and data sets.

A.1 Mean Errors And Negative Log Likelihood

This section shows the mean error and negative log likelihood results.

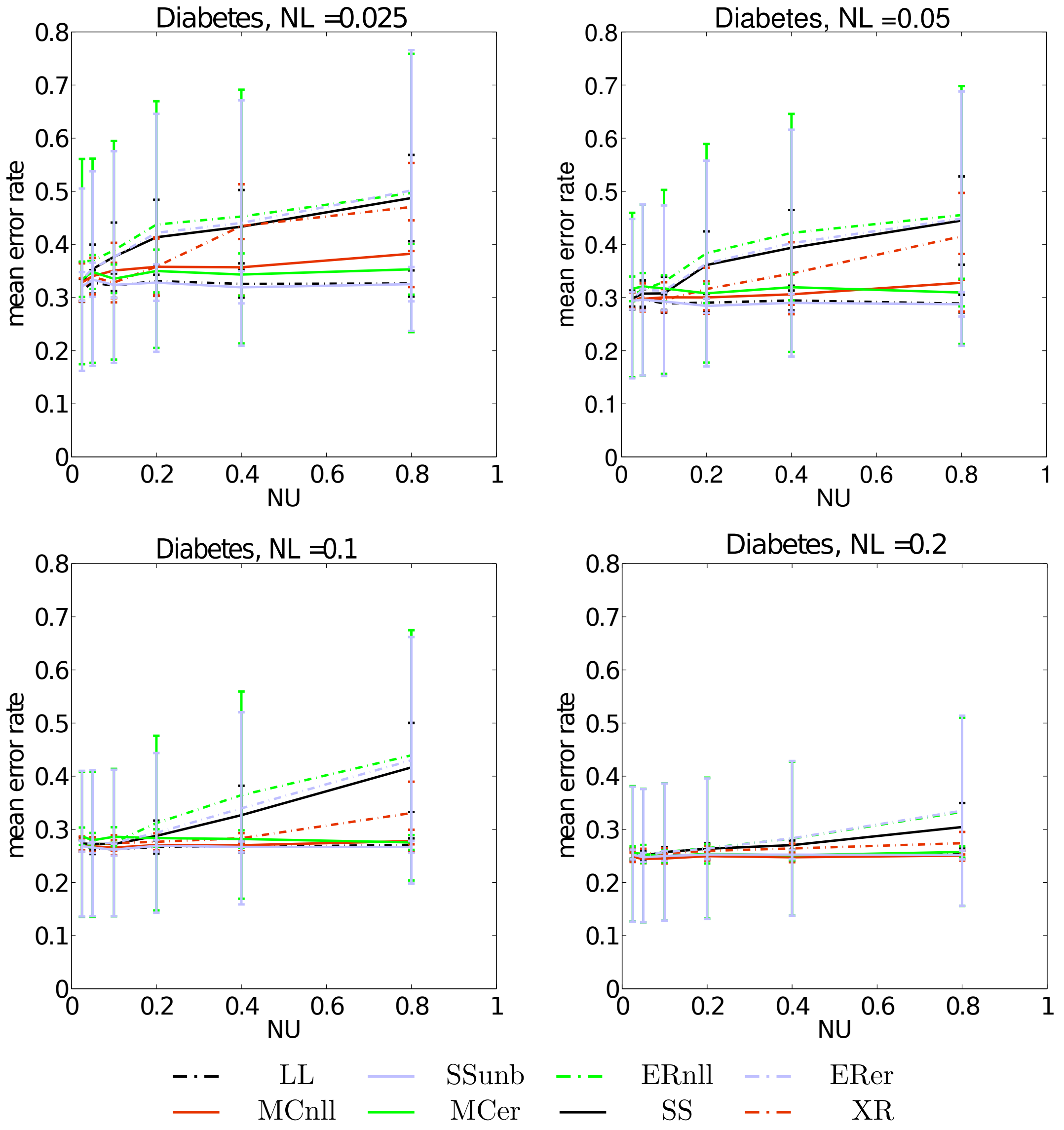

Figure 6: Mean error results of Diabetes data set

A.2 Proportion In Which Performance Improves

This section shows the proportion of test in which the error / negative log likelihood was improved by the addition of unlabelled data. That is, for every data set, the performance of the model was evaluated for each training scheme, and compared to the performance when only the labelled data was used. The frequency with which each semi-supervised scheme outperforms supervised learning was recorded and normalised. This gives an estimate of the probability that including unlabelled data will improve performance compared to supervised learning alone. A value close to 1 indicates reliable improvement when unlabelled data is added, whereas one close to 0 shows reliable worsening of results.

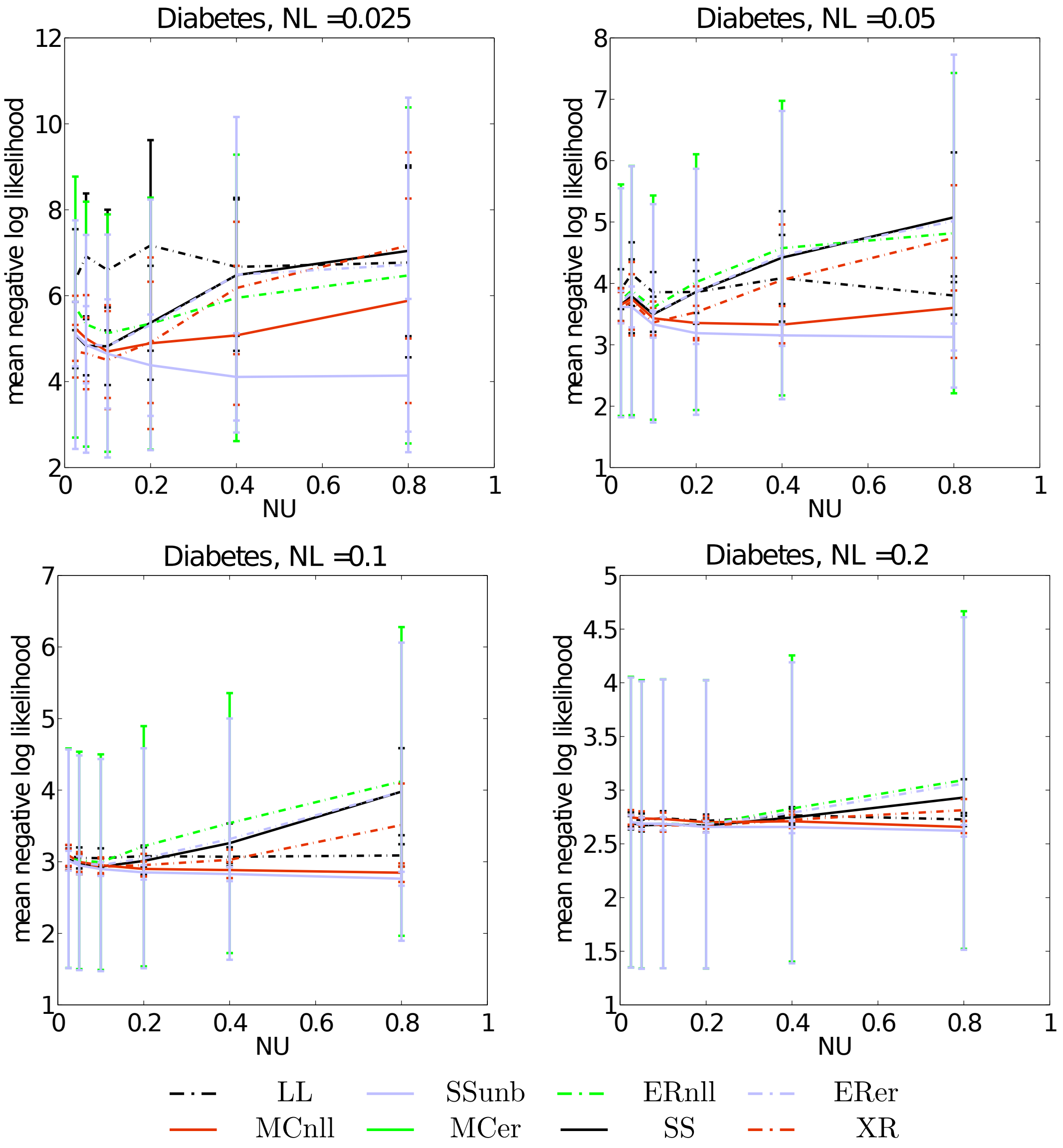

Figure 7: Mean log likelihood results of Diabetes data set

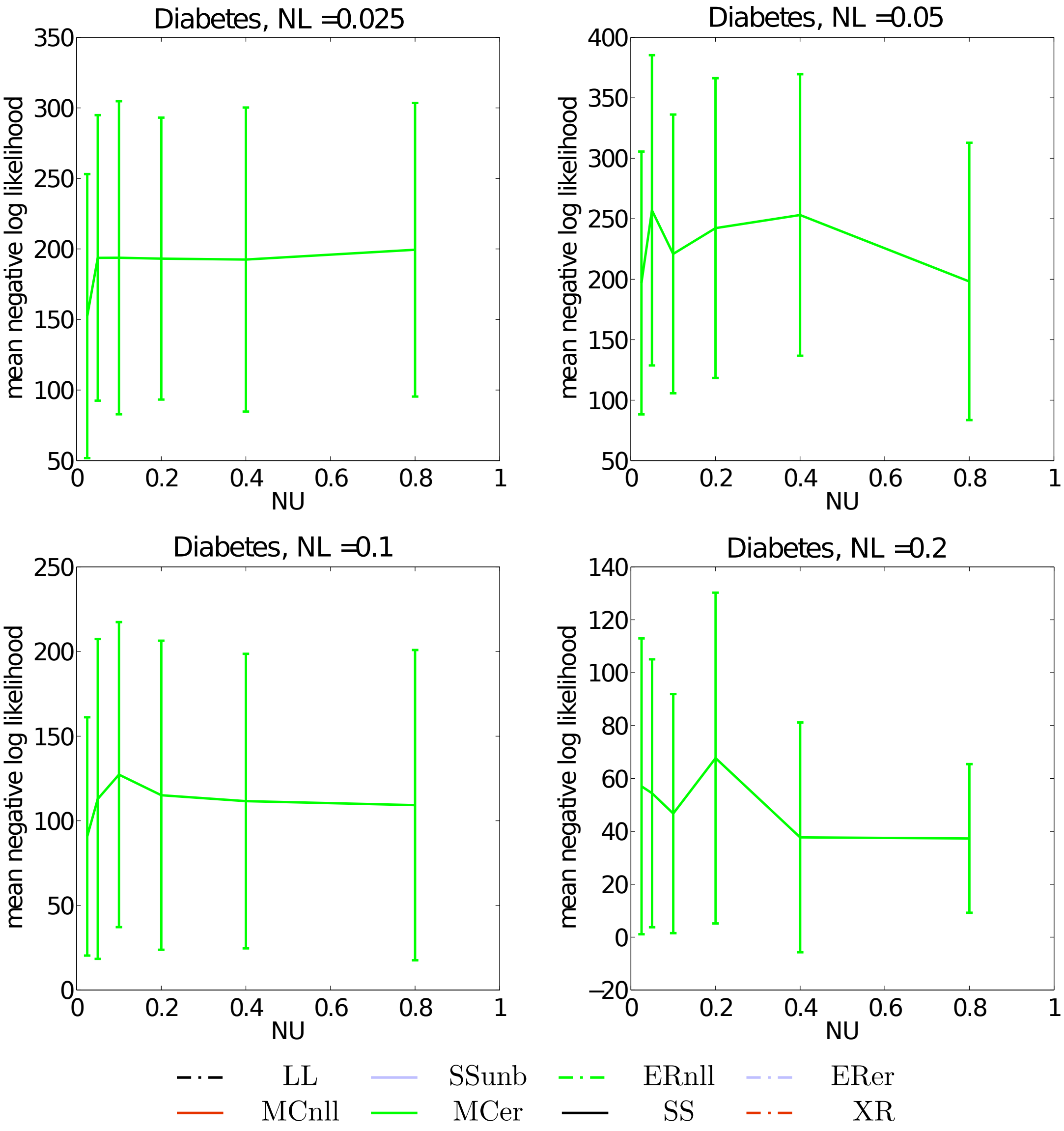

Figure 8: Mean log likelihood results of Diabetes data set

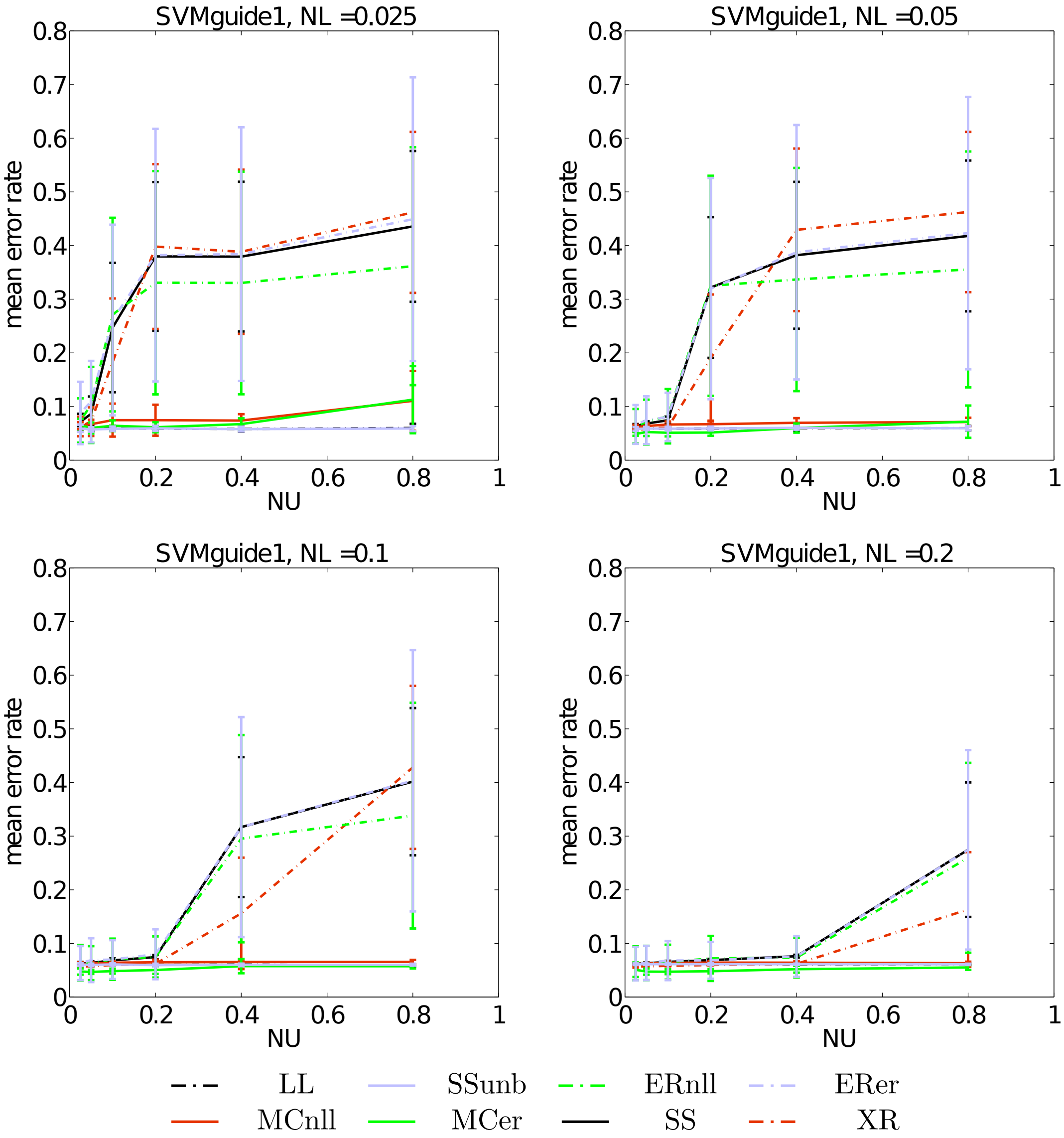

Figure 9: Mean error results of the SVMguide data set

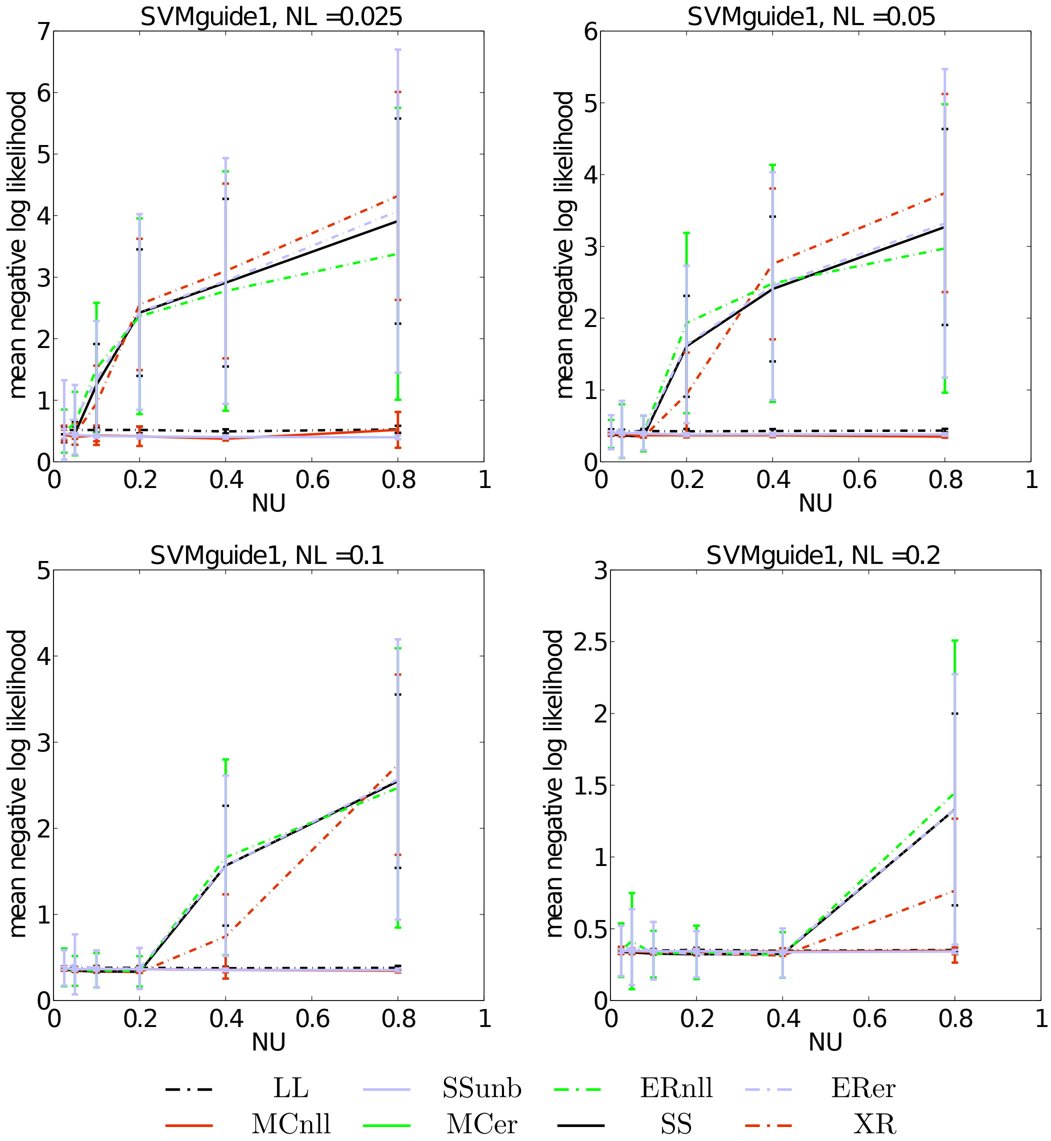

Figure 10: Mean log likelihood results of the SVMguide data set

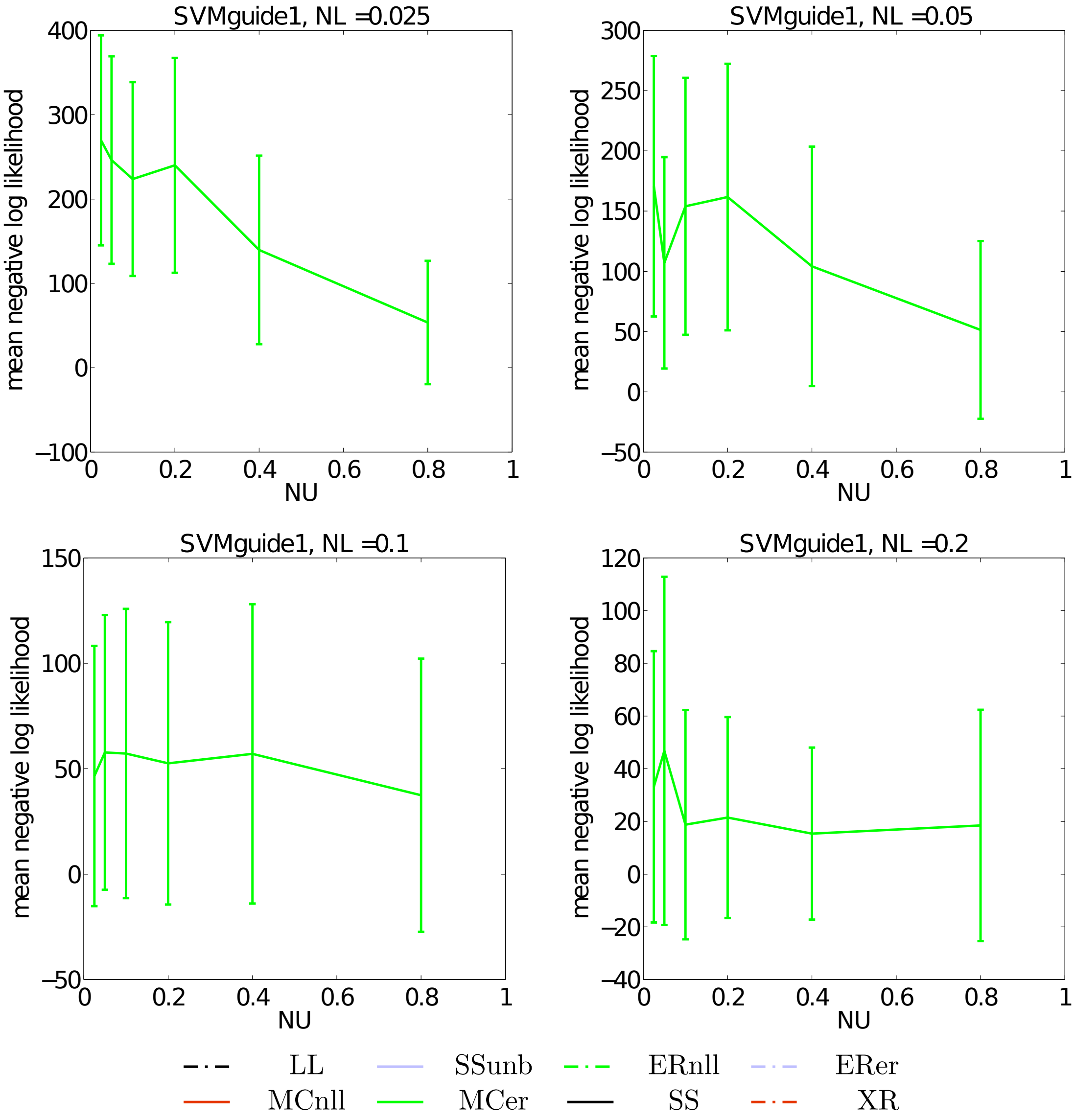

Figure 11: Mean log likelihood results of the SVMguide data set

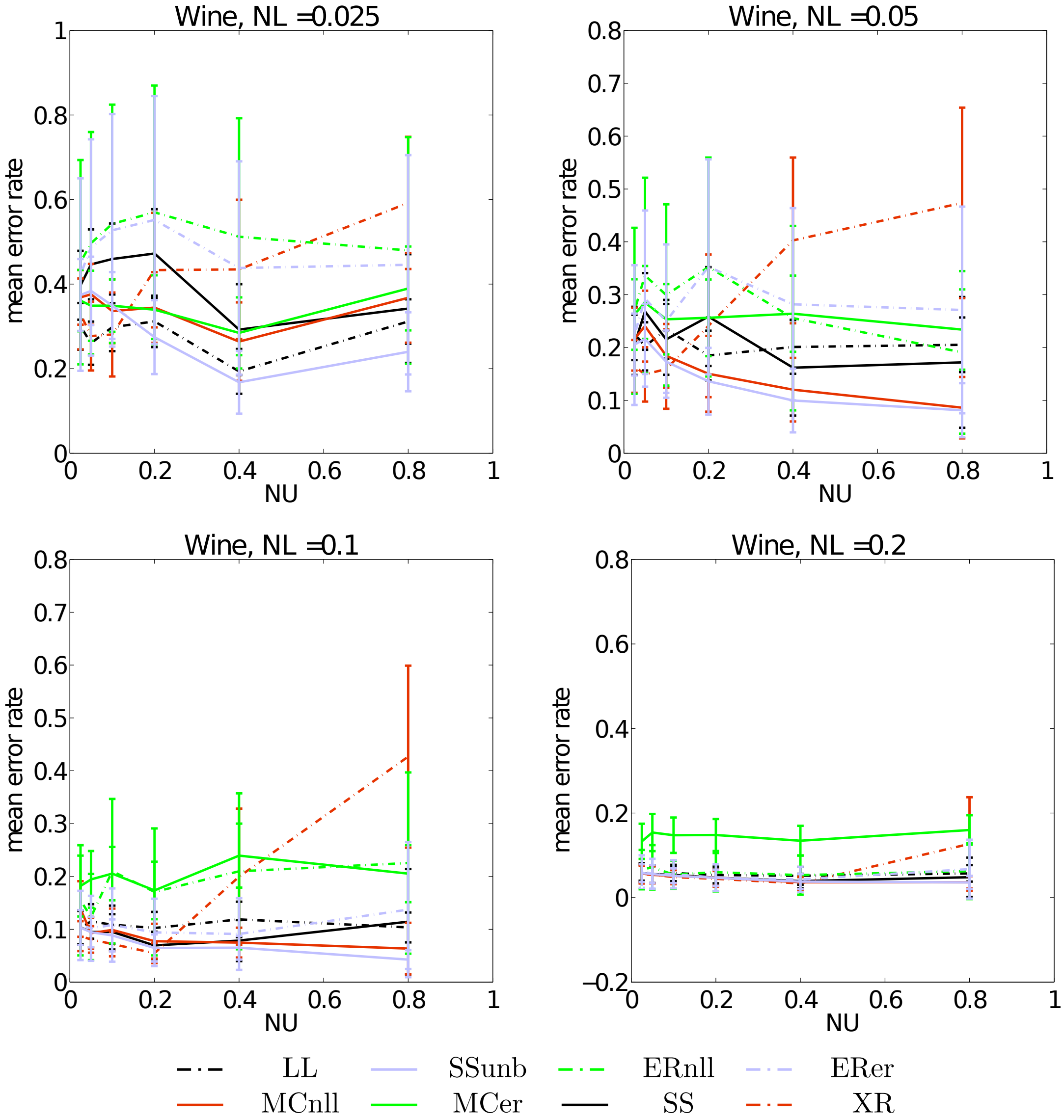

Figure 12: Mean error results of the Wine data set

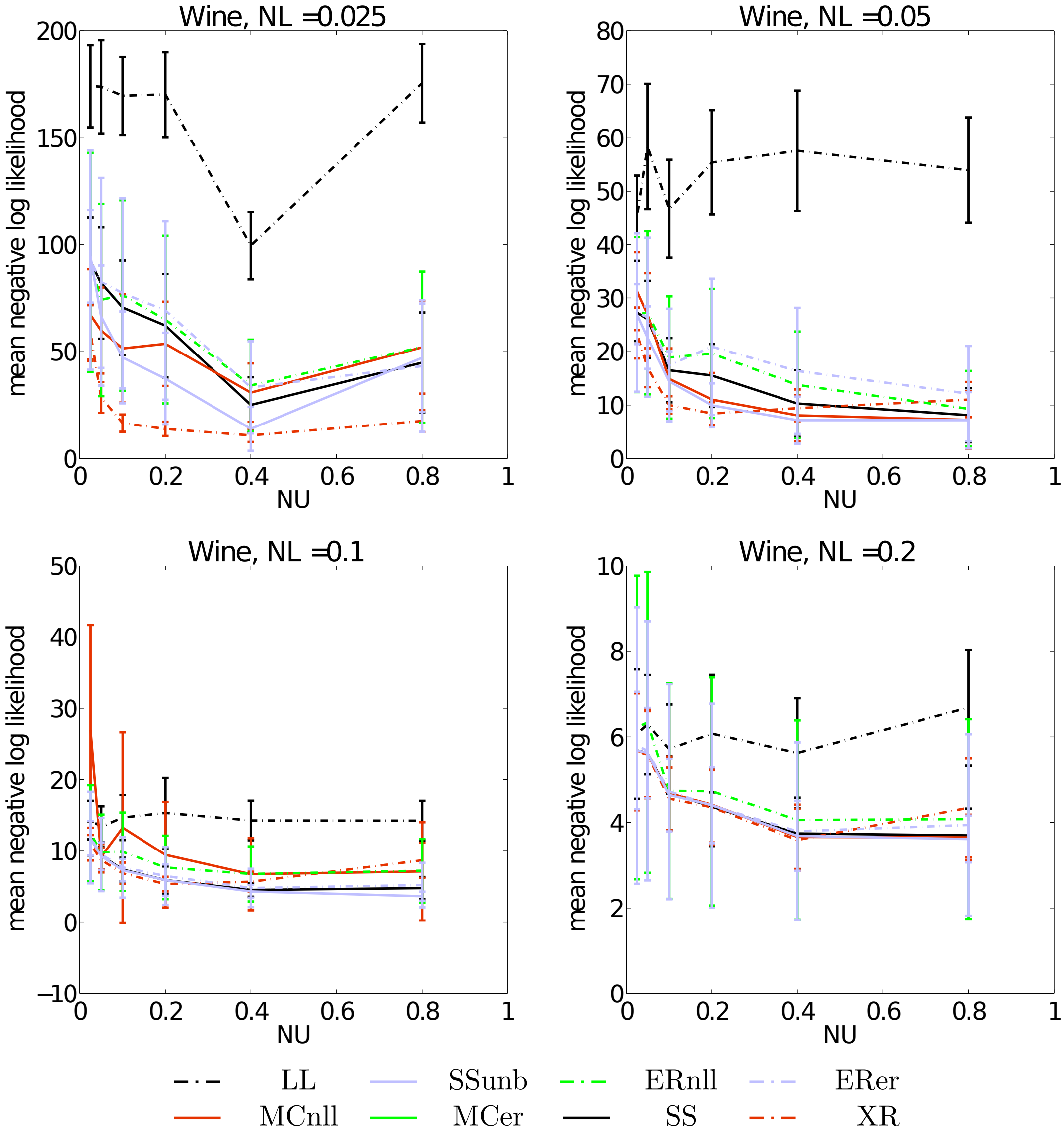

Figure 13: Mean log likelihood results of the Wine data set

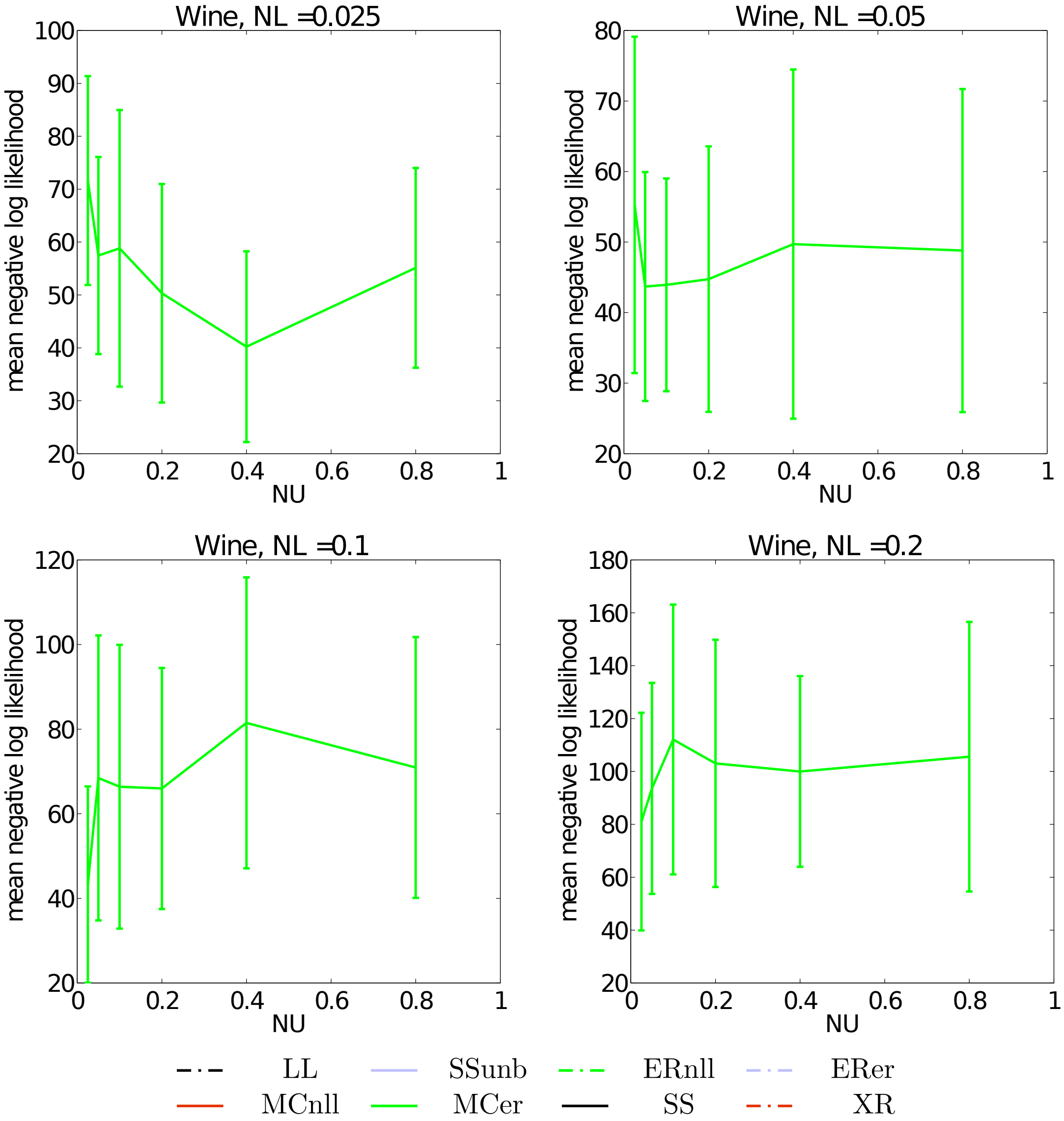

Figure 14: Mean log likelihood results of the Wine data set

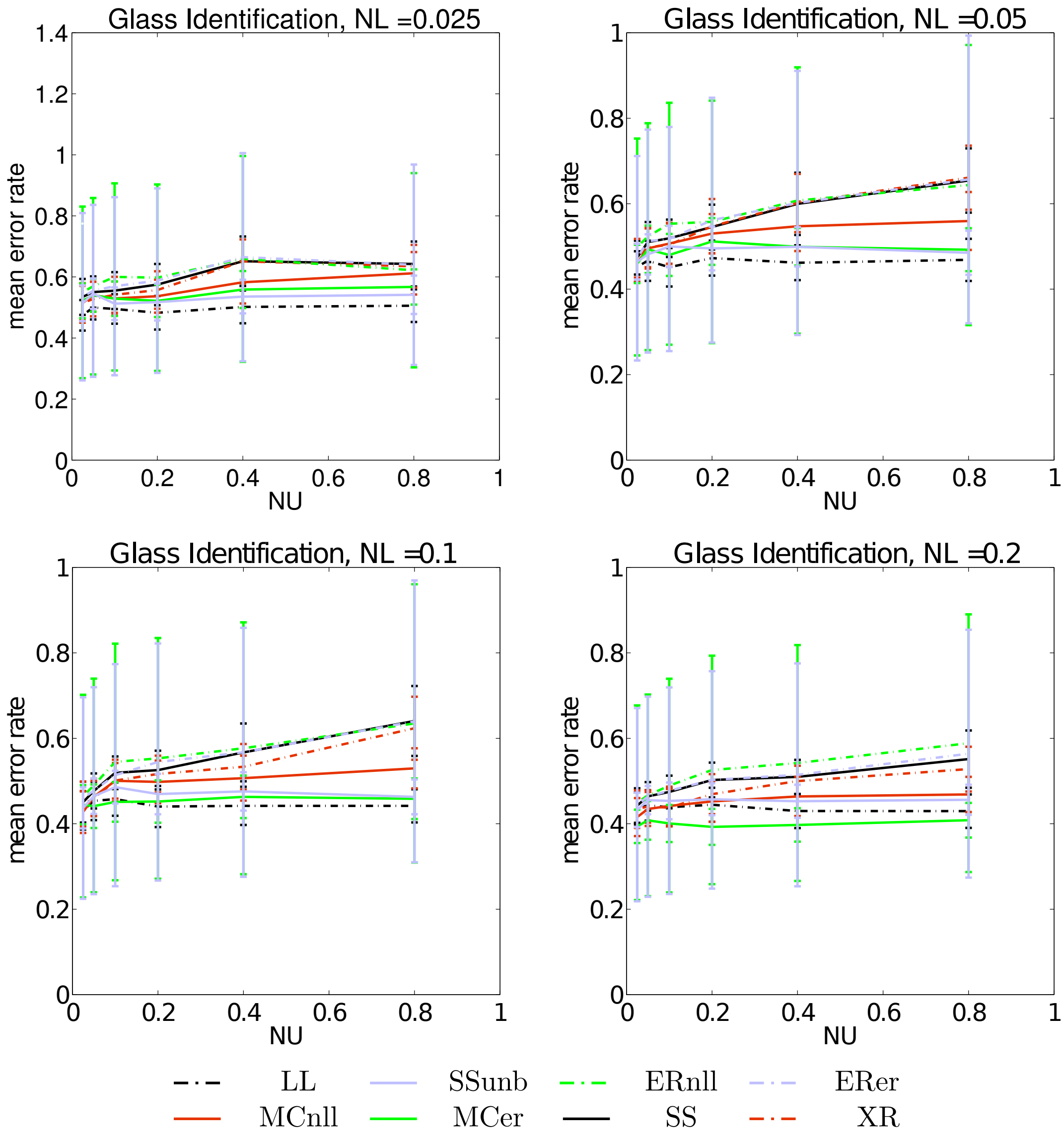

Figure 15: Mean error results of the Glass data set

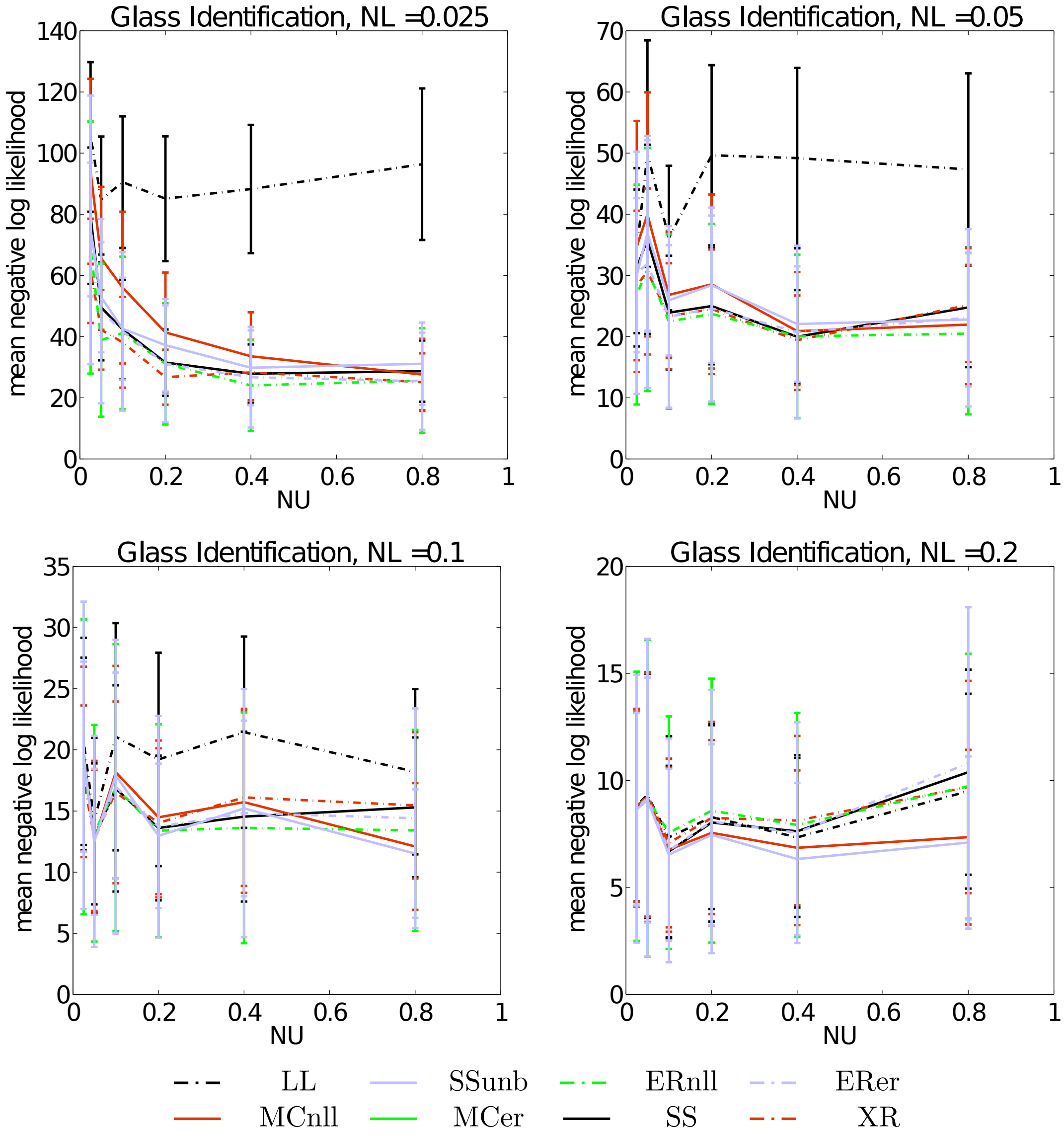

Figure 16: Mean log likelihood results of the Glass data set

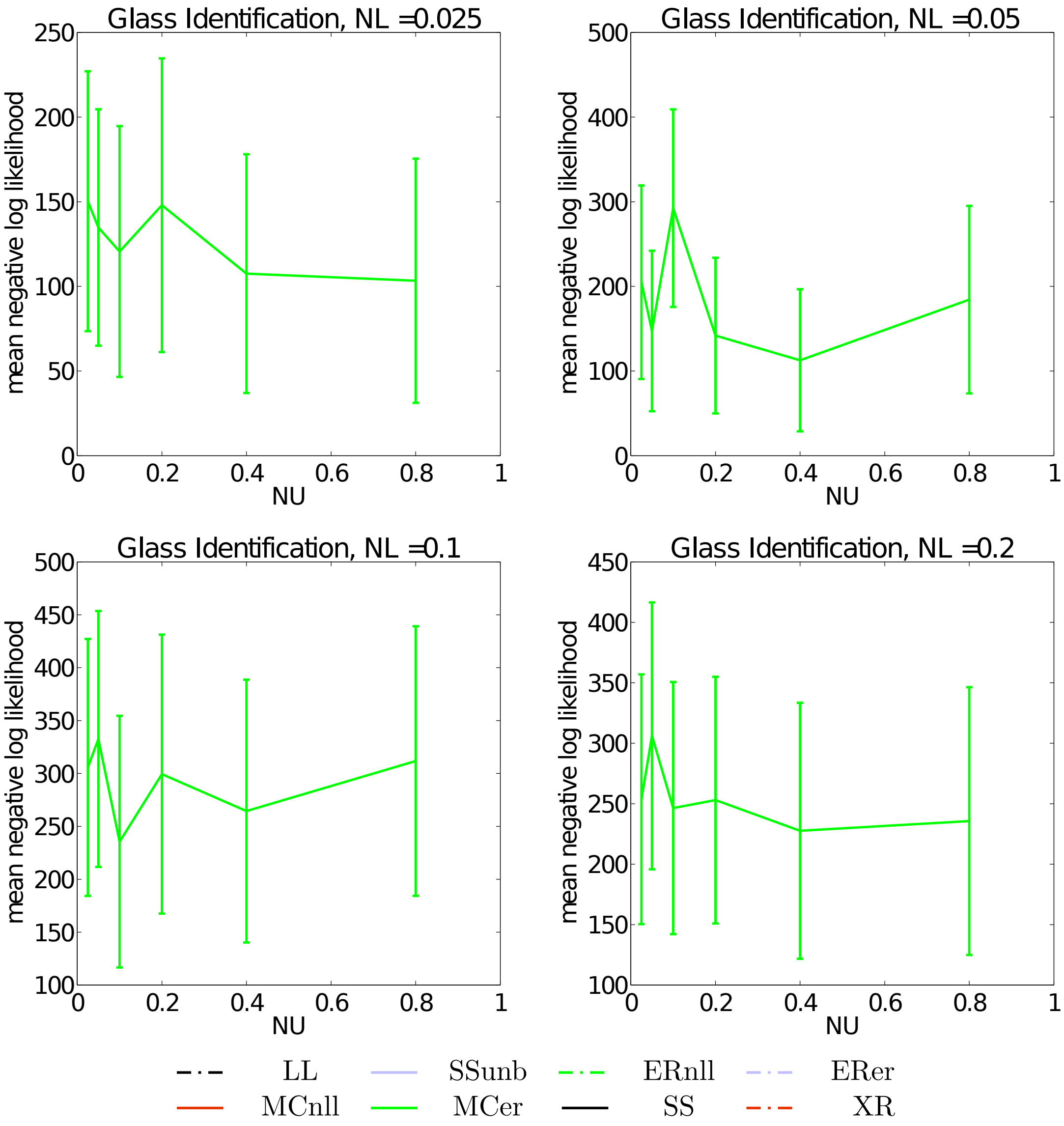

Figure 17: Mean log likelihood results of the Glass data set

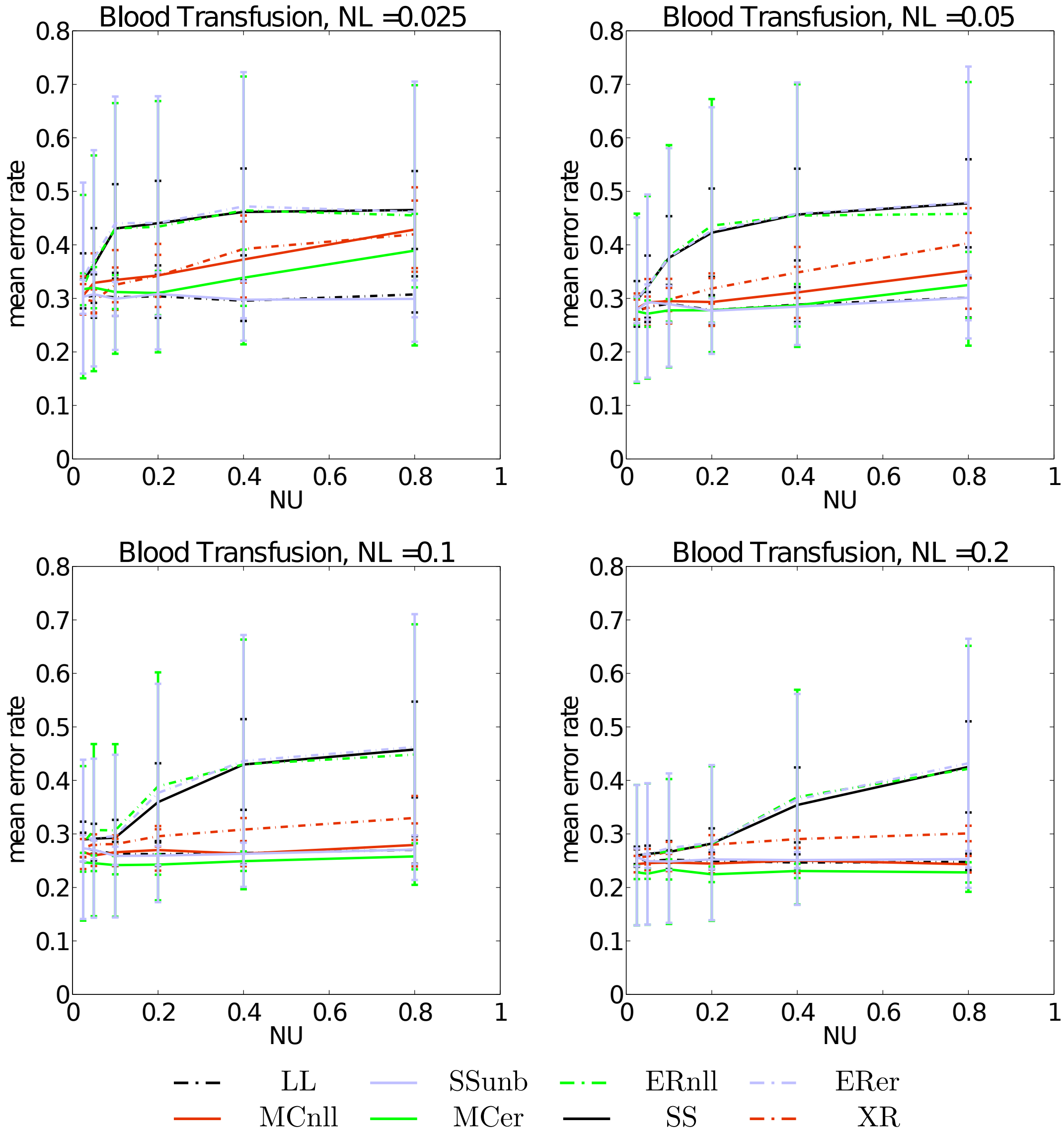

Figure 18: Mean error results of the Blood data set

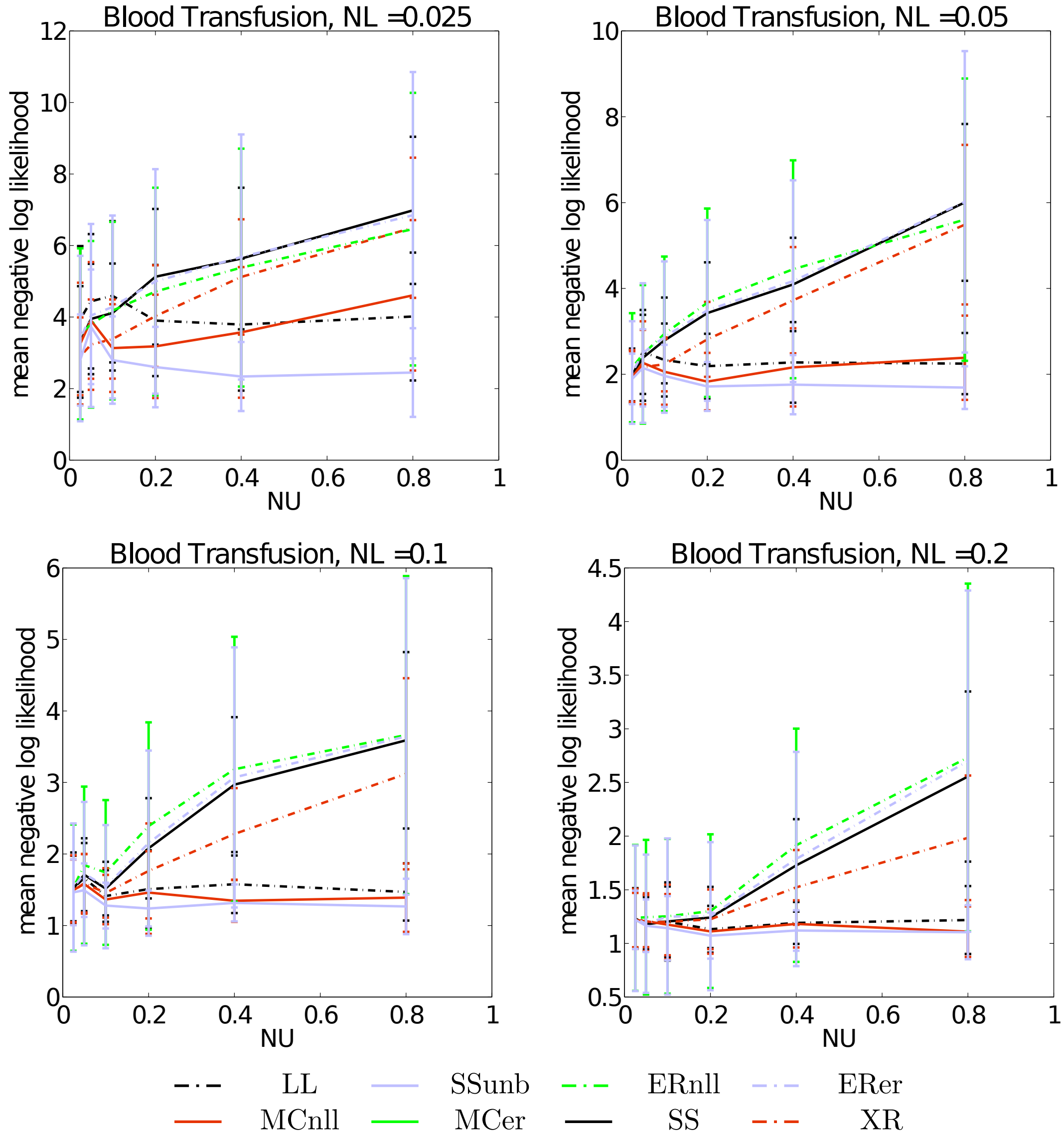

Figure 19: Mean log likelihood results of the Blood data set

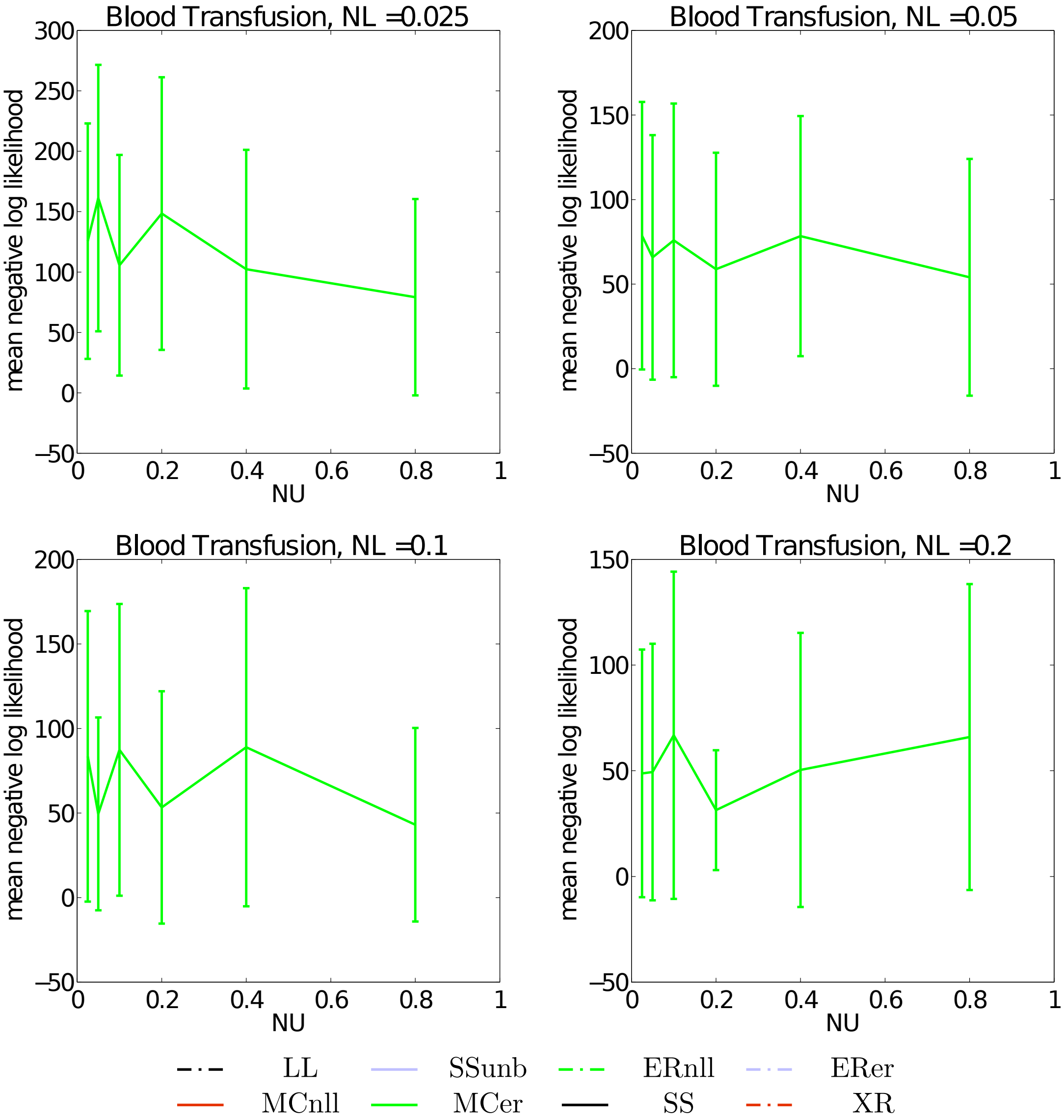

Figure 20: Mean log likelihood results of the Blood data set

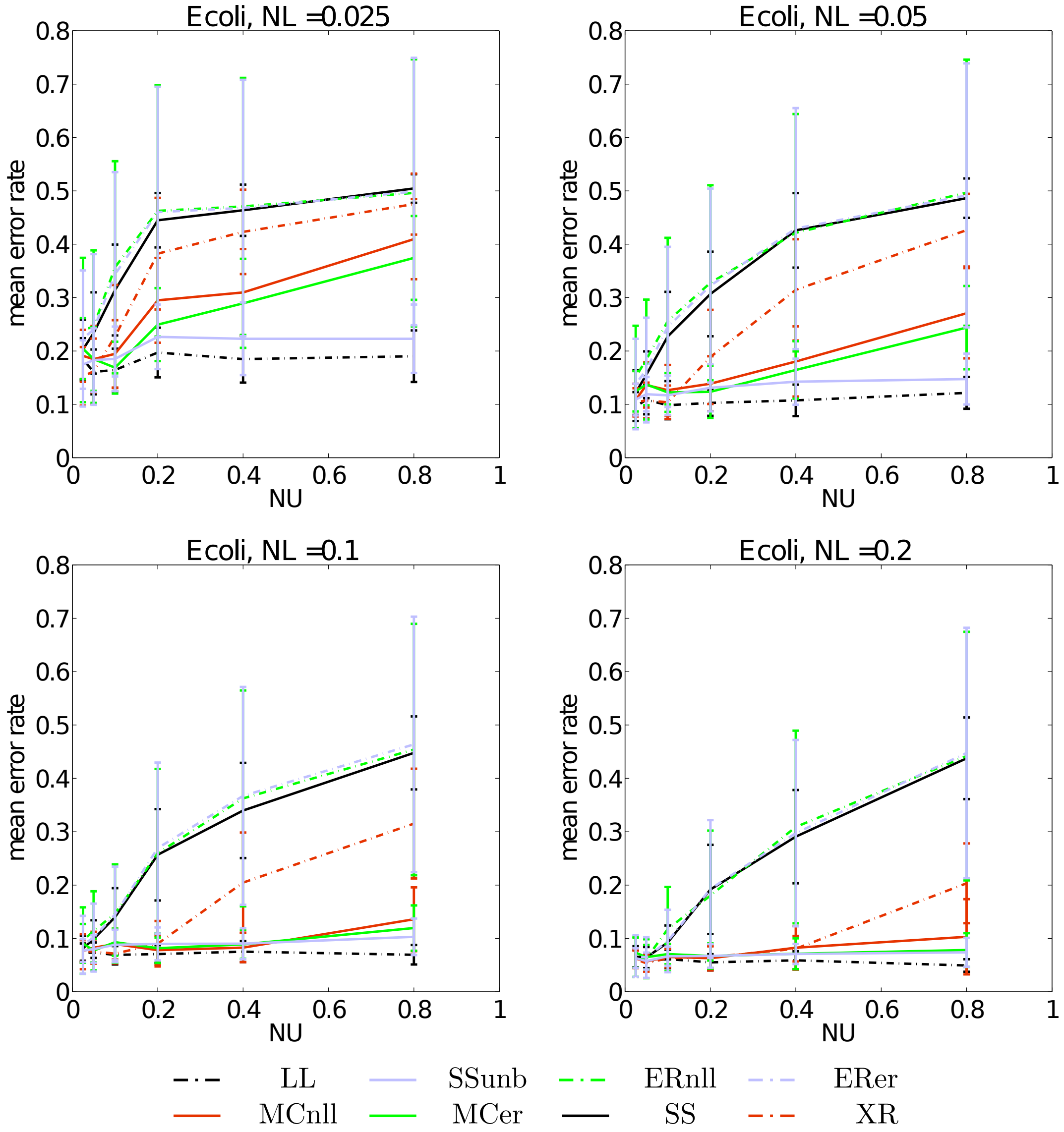

Figure 21: Mean error results of the Ecoil data set

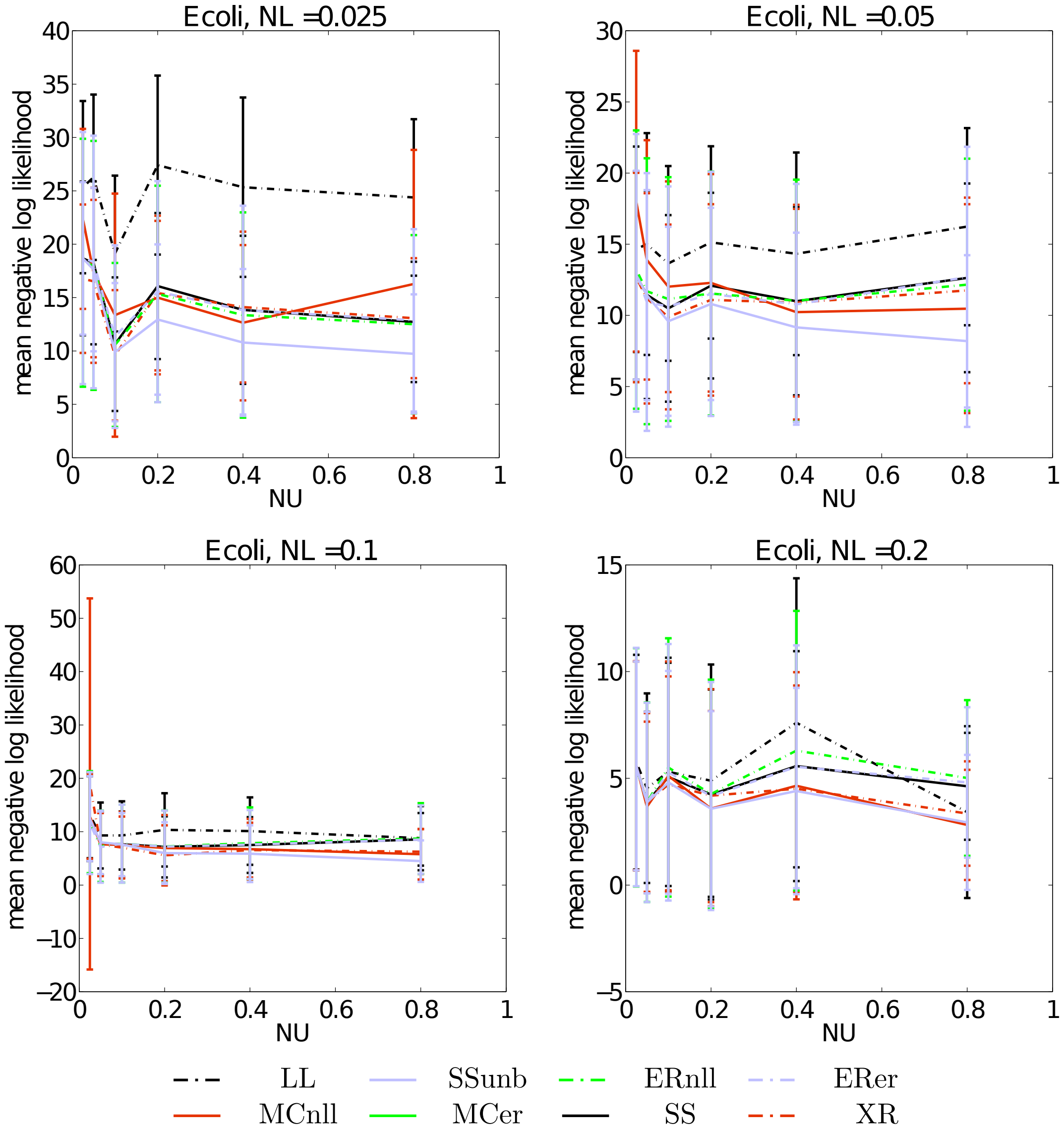

Figure 22: Mean log likelihood results of the Ecoil data set

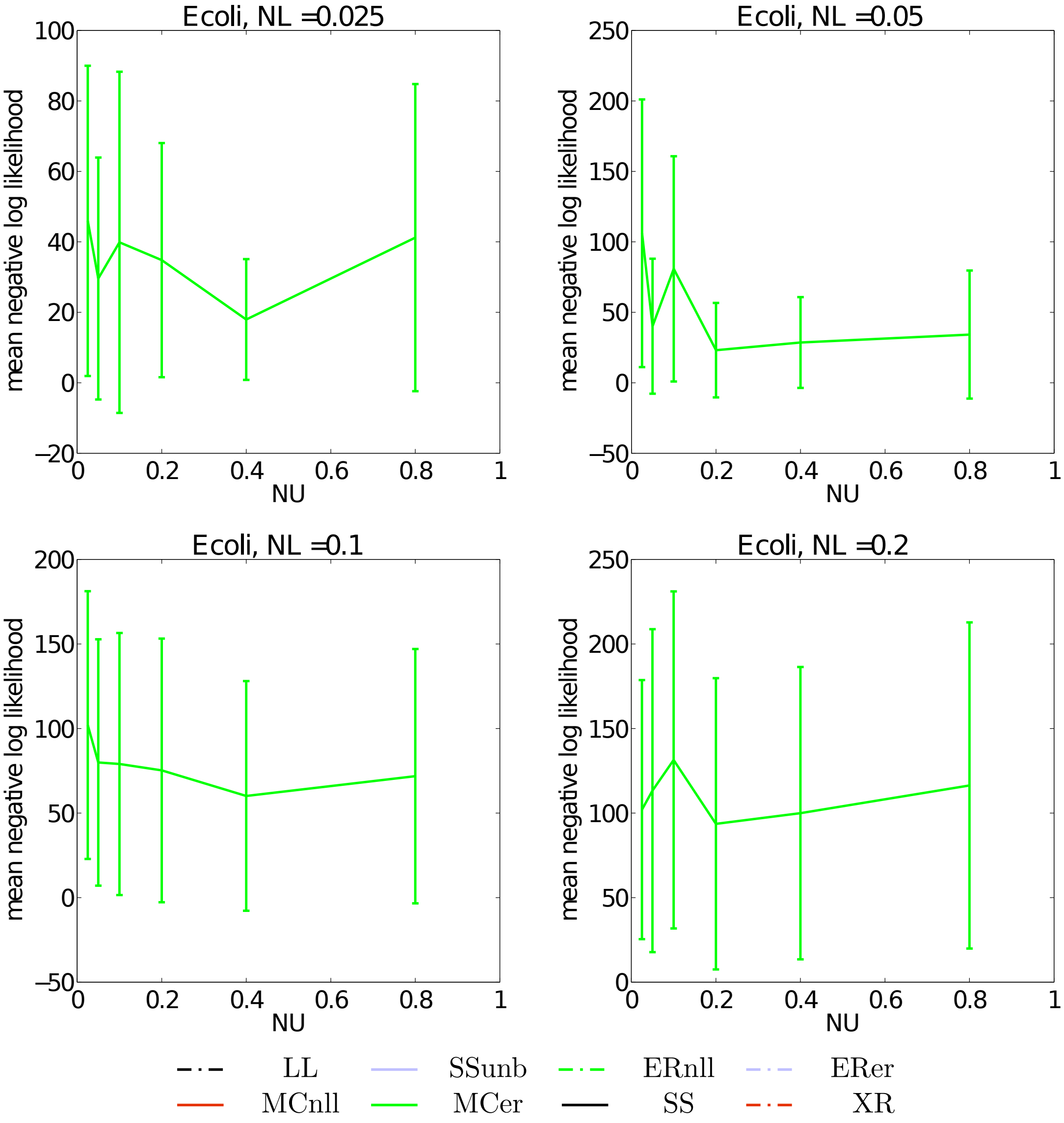

Figure 23: Mean log likelihood results of the Ecoli data set

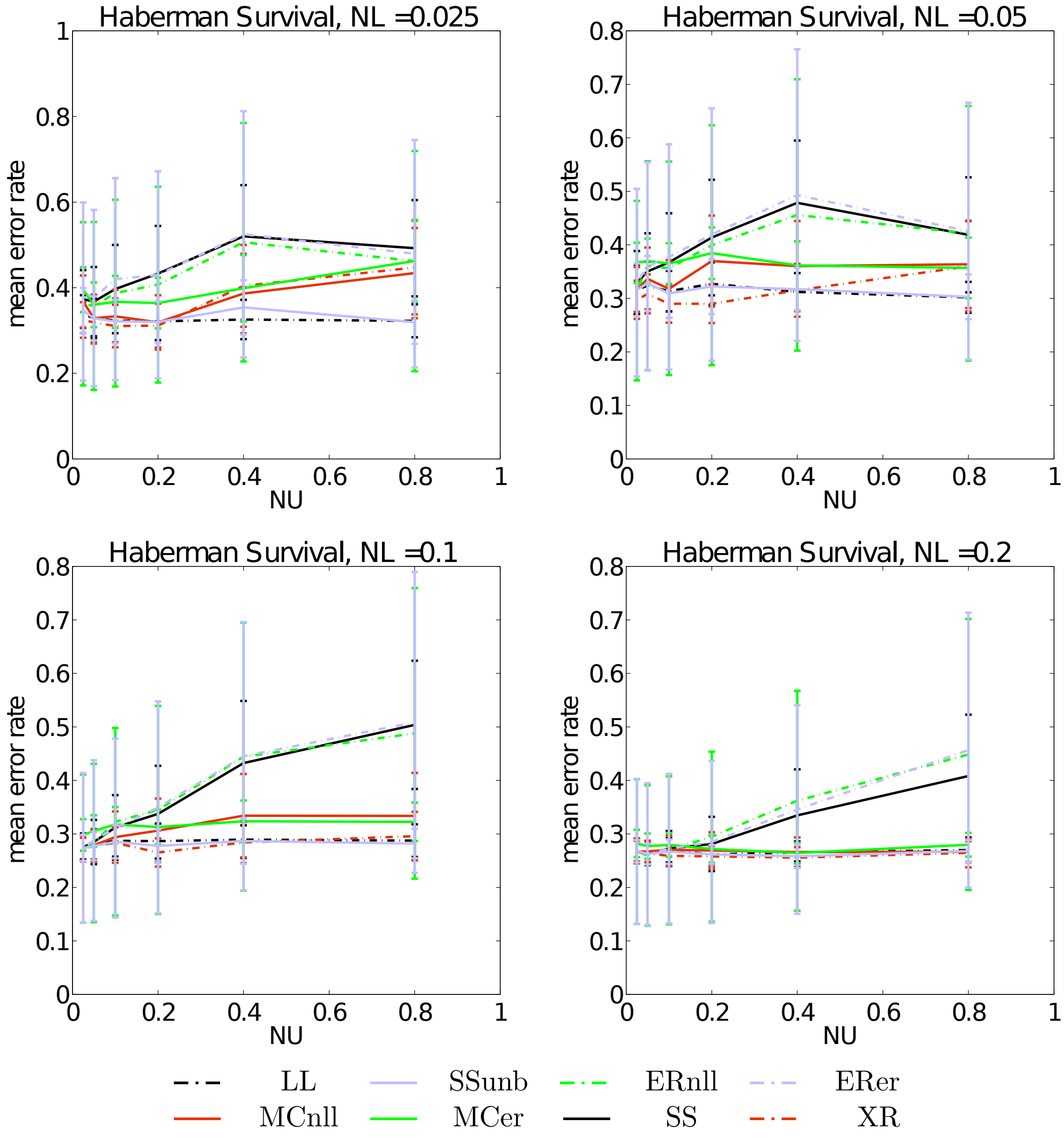

Figure 24: Mean error results of the Haberman data set

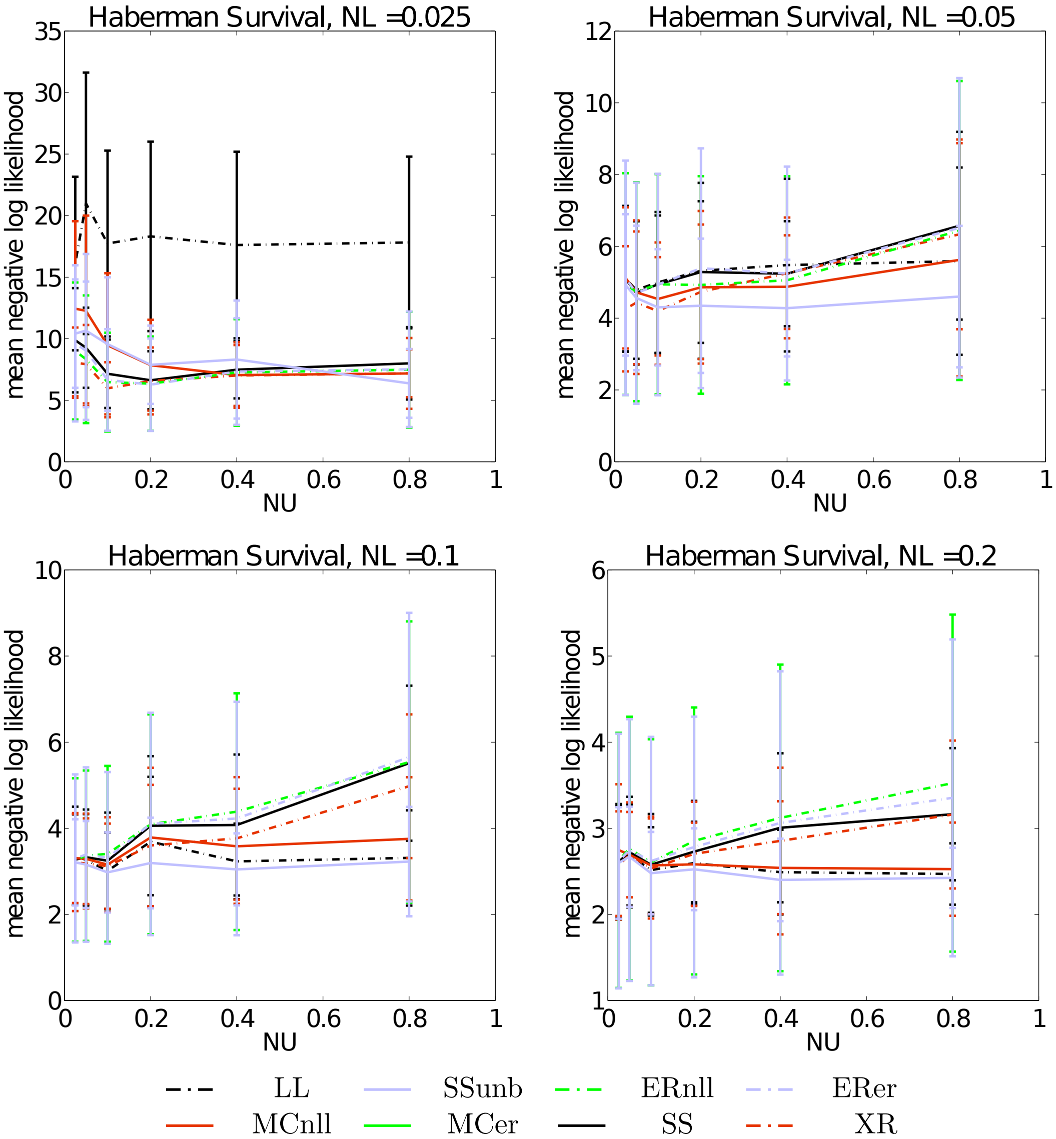

Figure 25: Mean log likelihood results of the Haberman data set

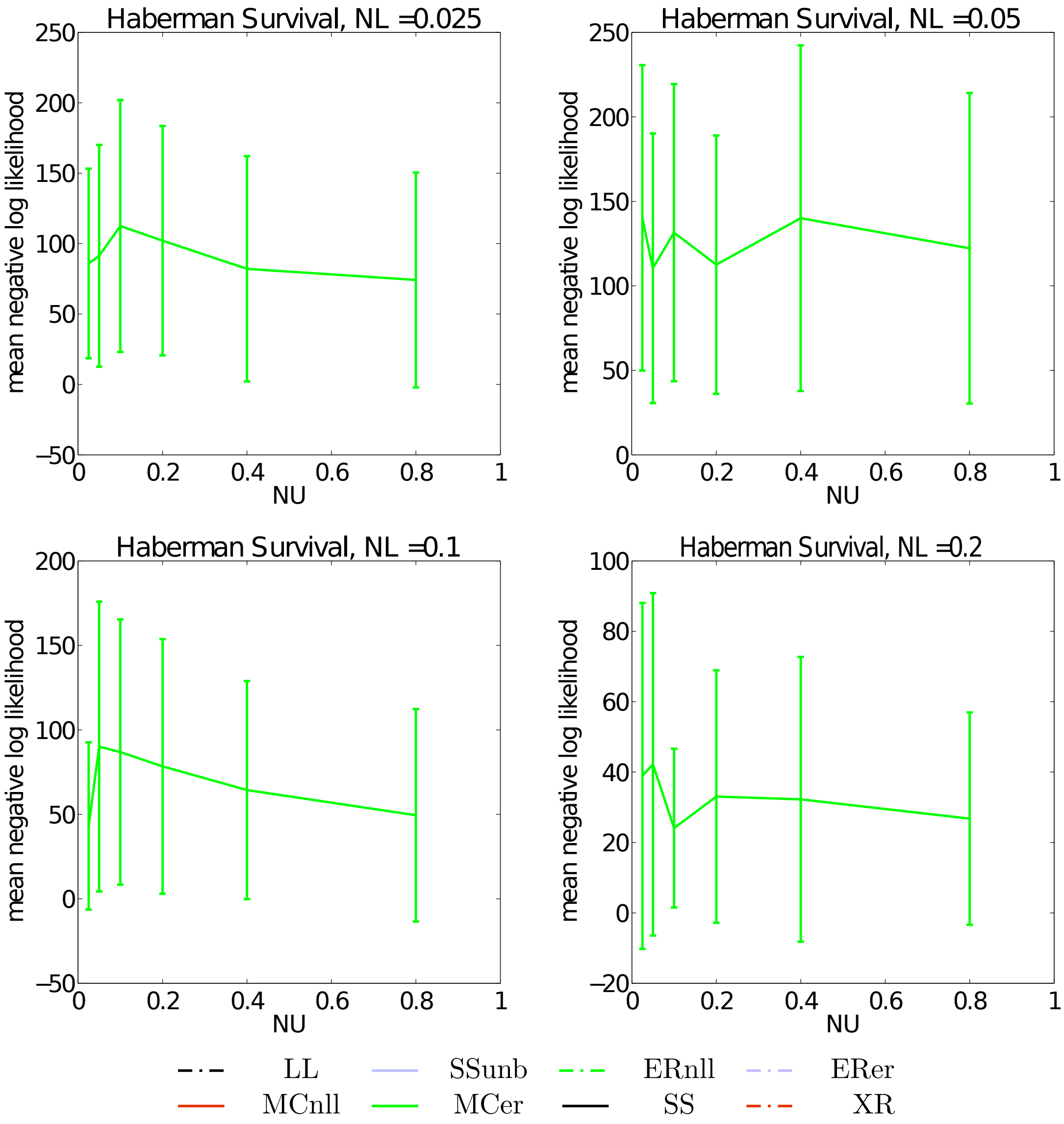

Figure 26: Mean log likelihood results of the Haberman data set

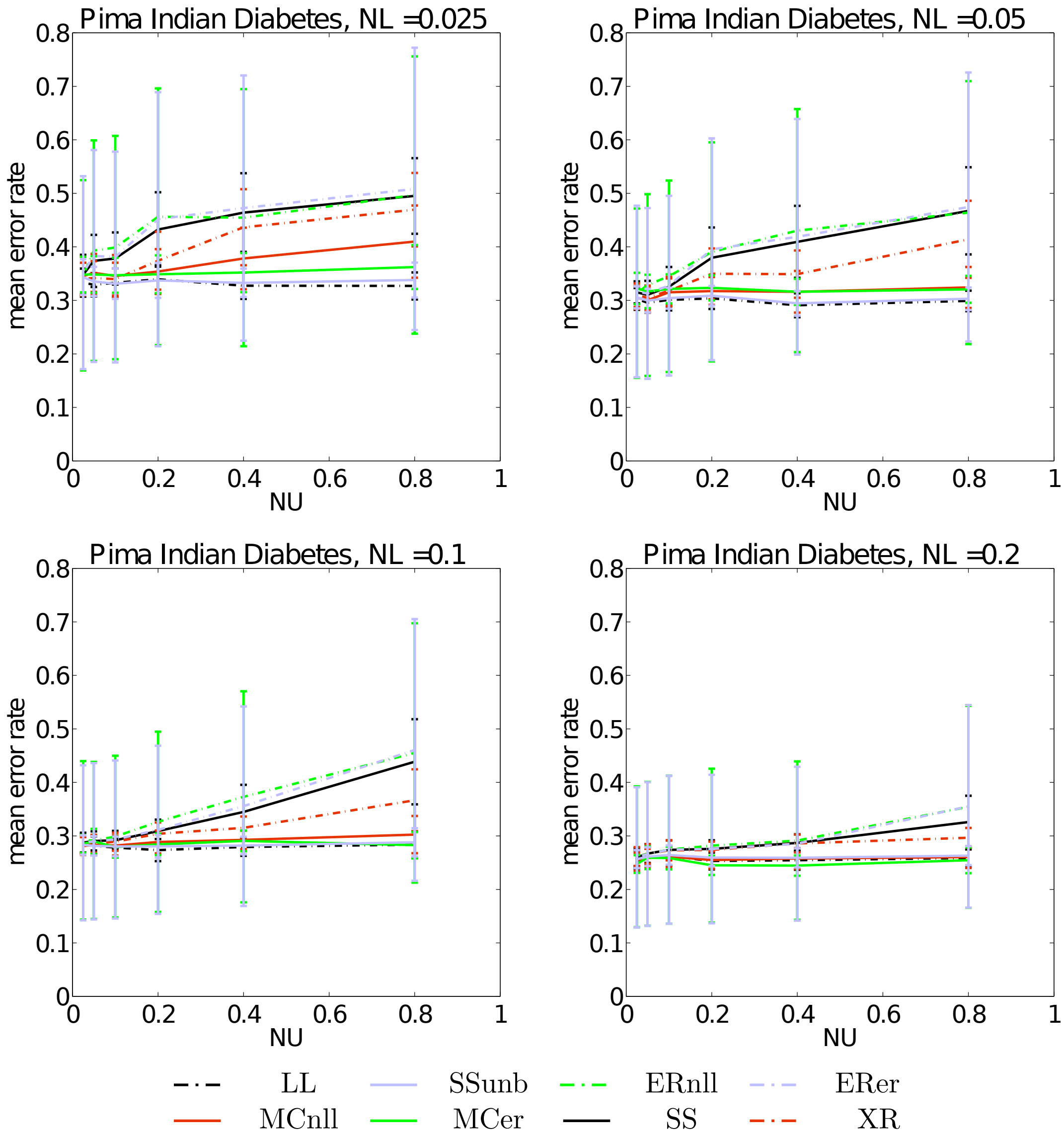

Figure 27: Mean error results of the Pima Indian data set

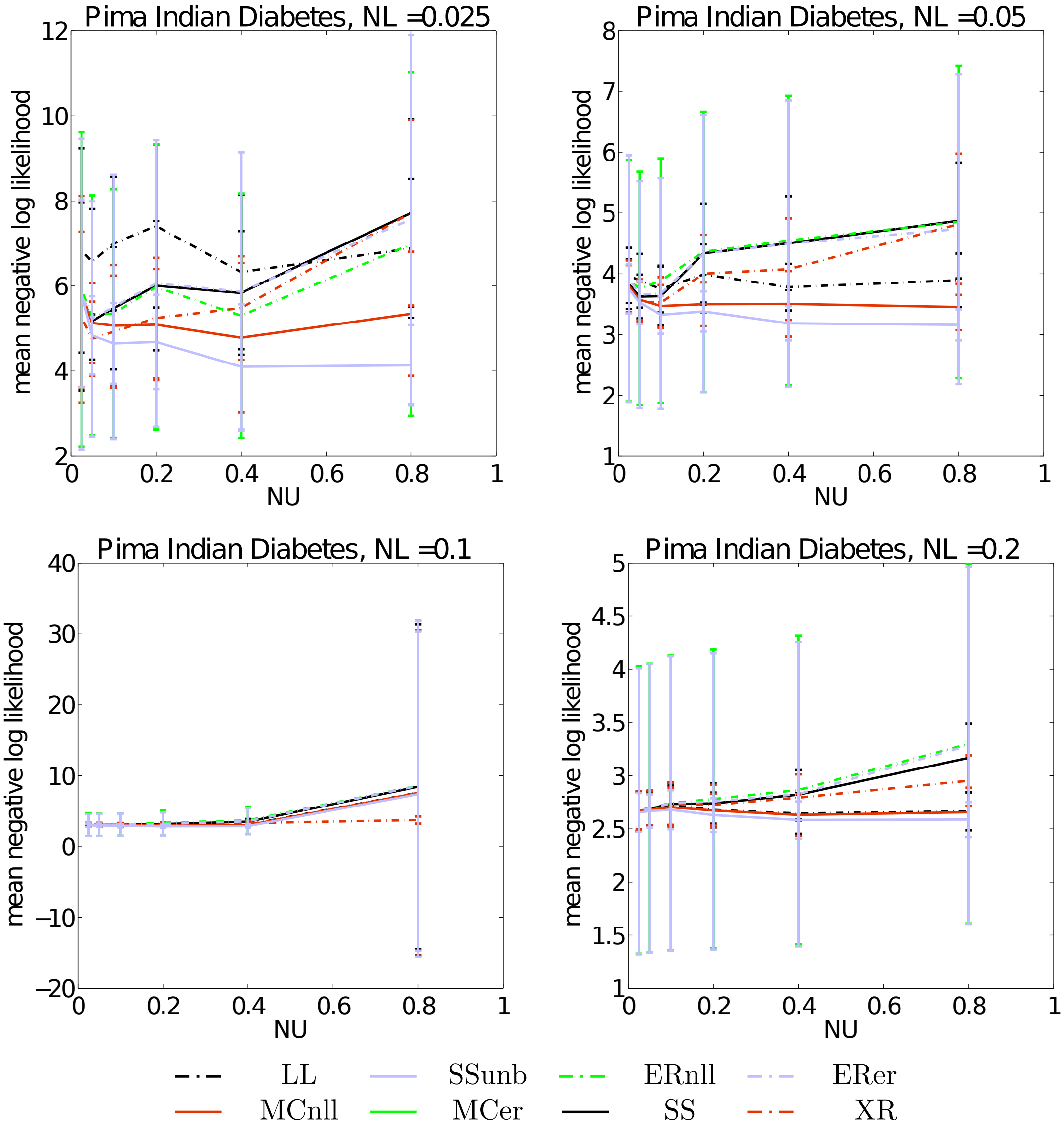

Figure 28: Mean log likelihood results of the Pima Indian data set

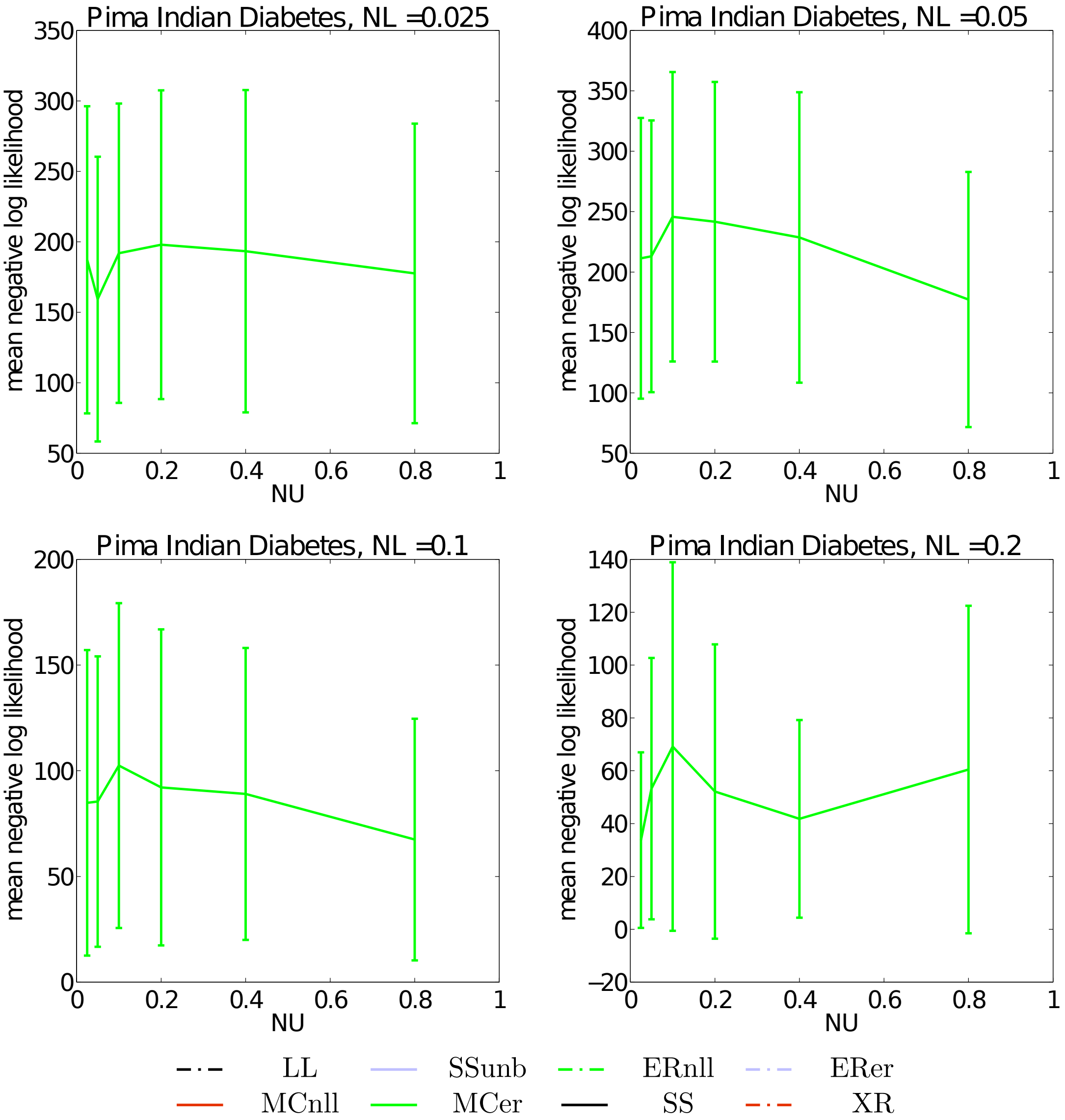

Figure 29: Mean log likelihood results of the Pima Indian data set

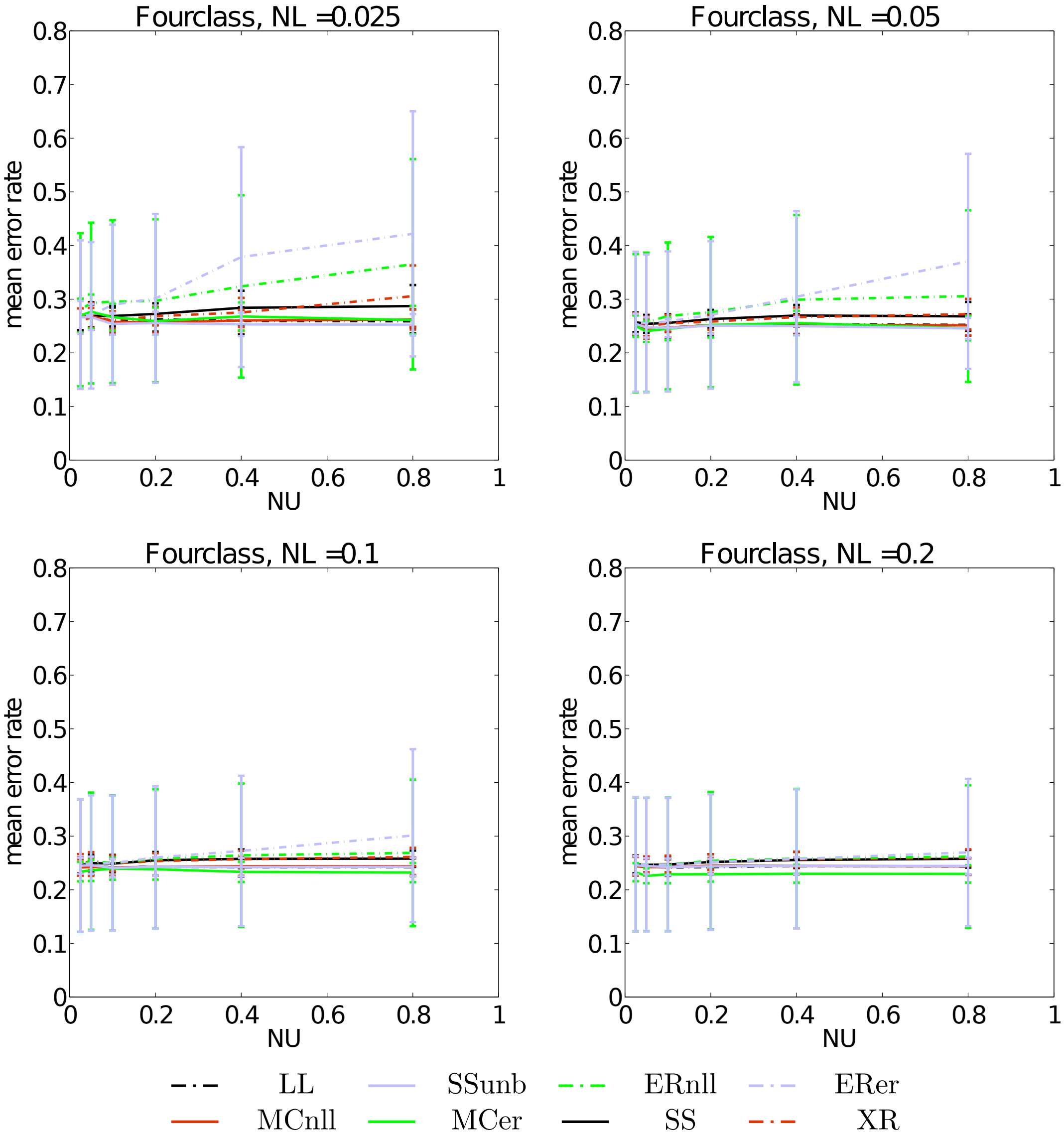

Figure 30: Mean error results of the Fourclass data set

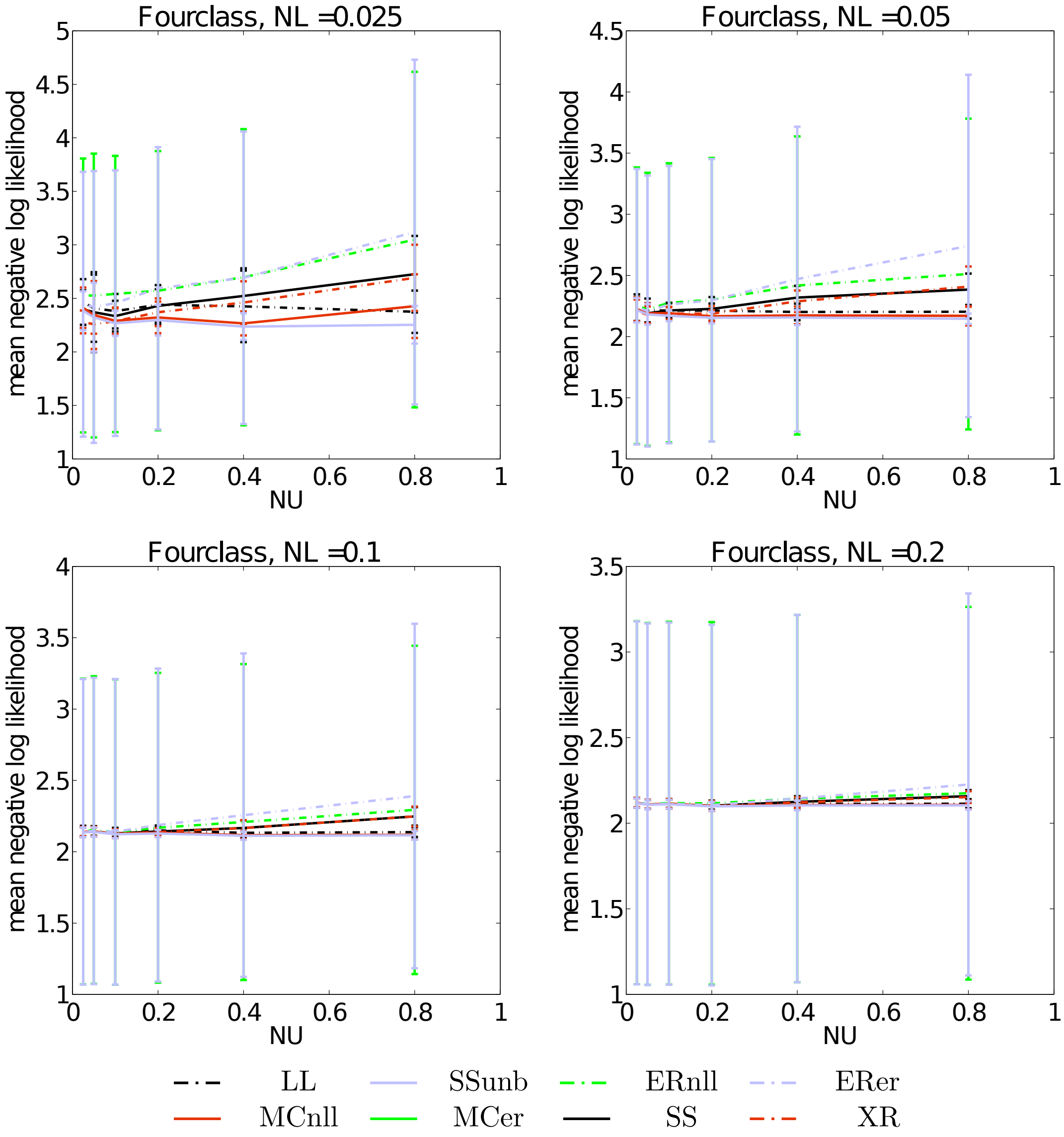

Figure 31: Mean log likelihood results of the Fourclass data set

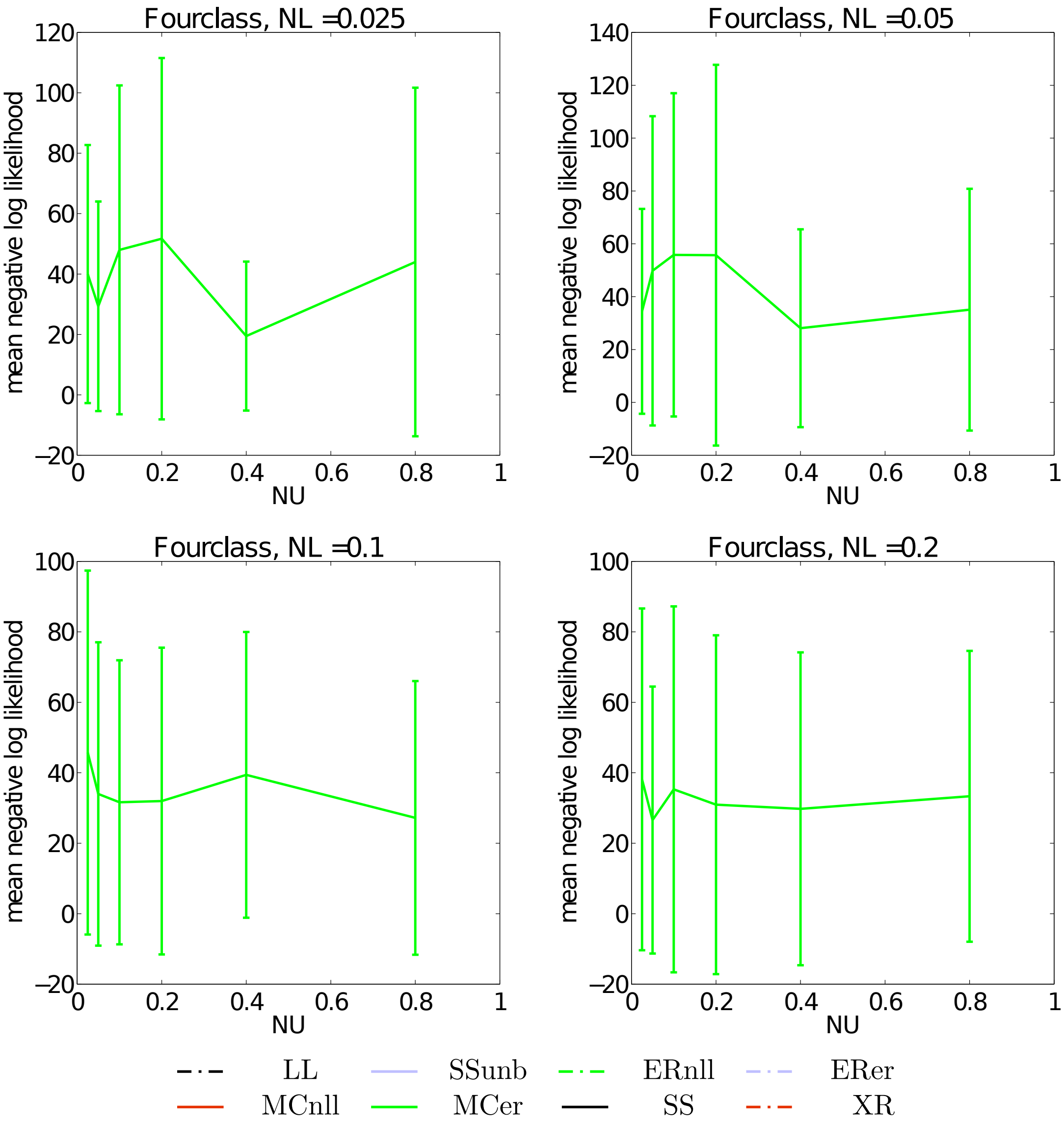

Figure 32: Mean log likelihood results of the Fourclass data set

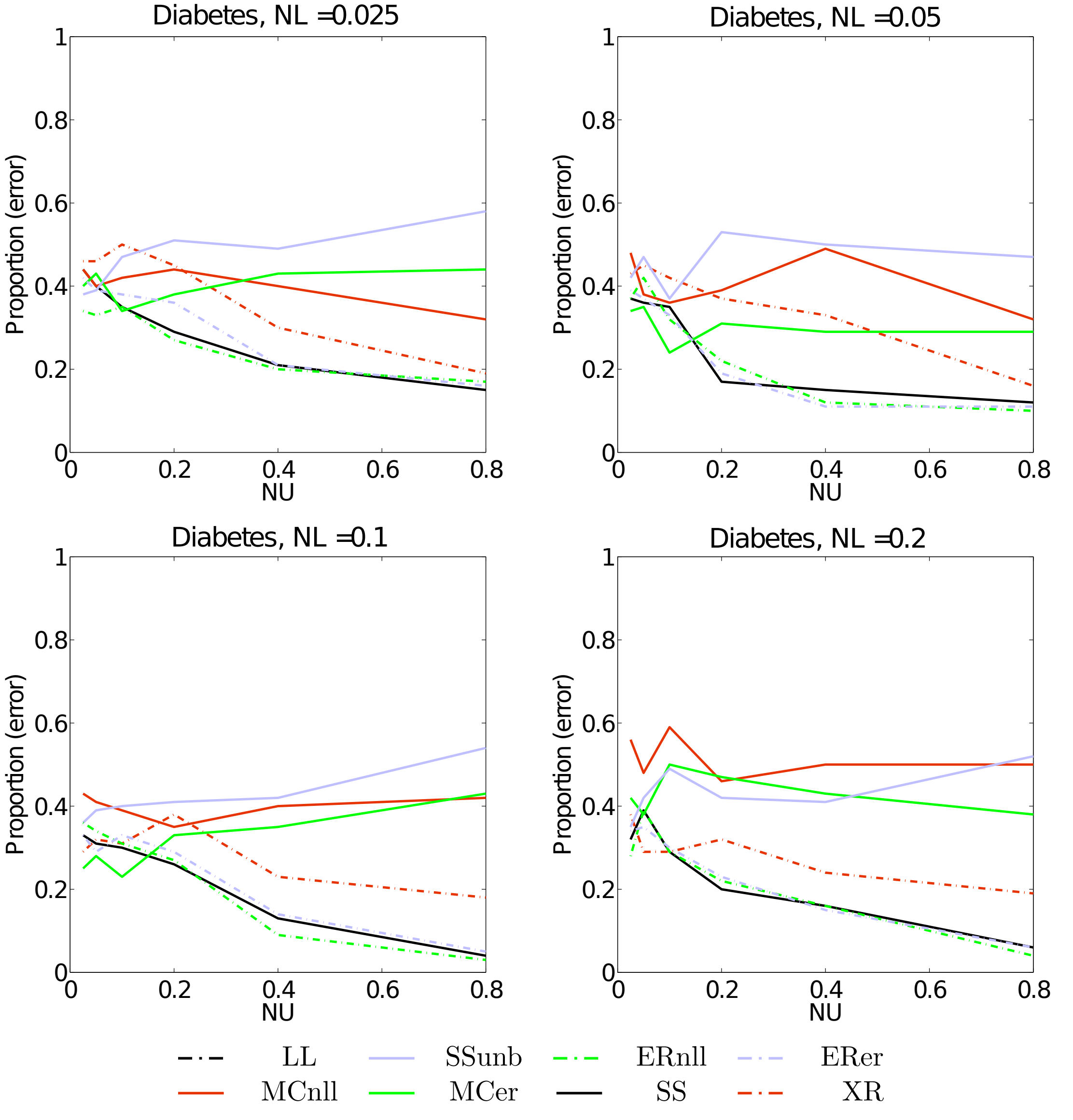

Figure 33: Chance of error rate improvement VS labelled data alone - Diabetes data set

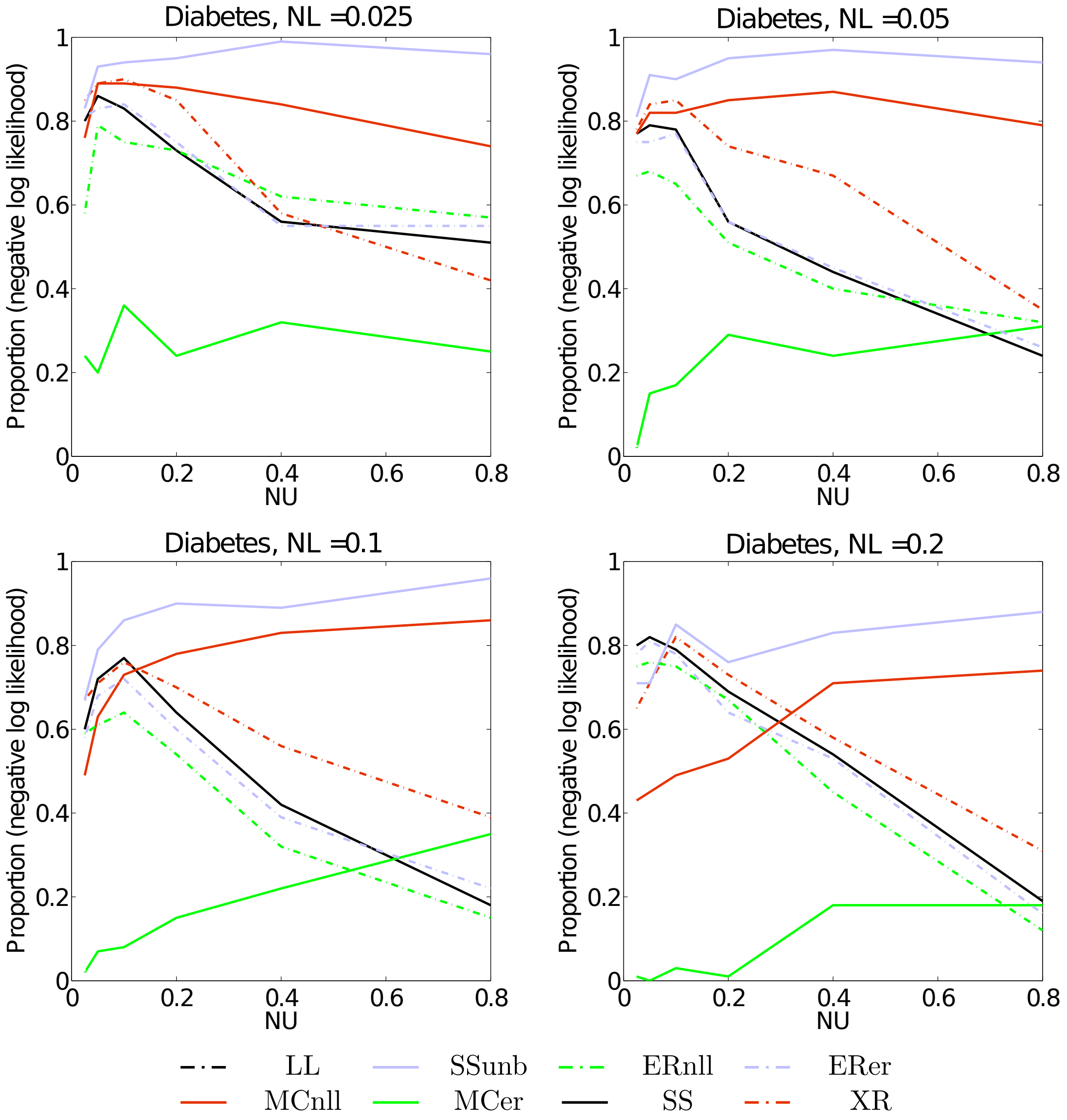

Figure 34: Chance of log likelihood improvement VS labelled data alone - Diabetes data set

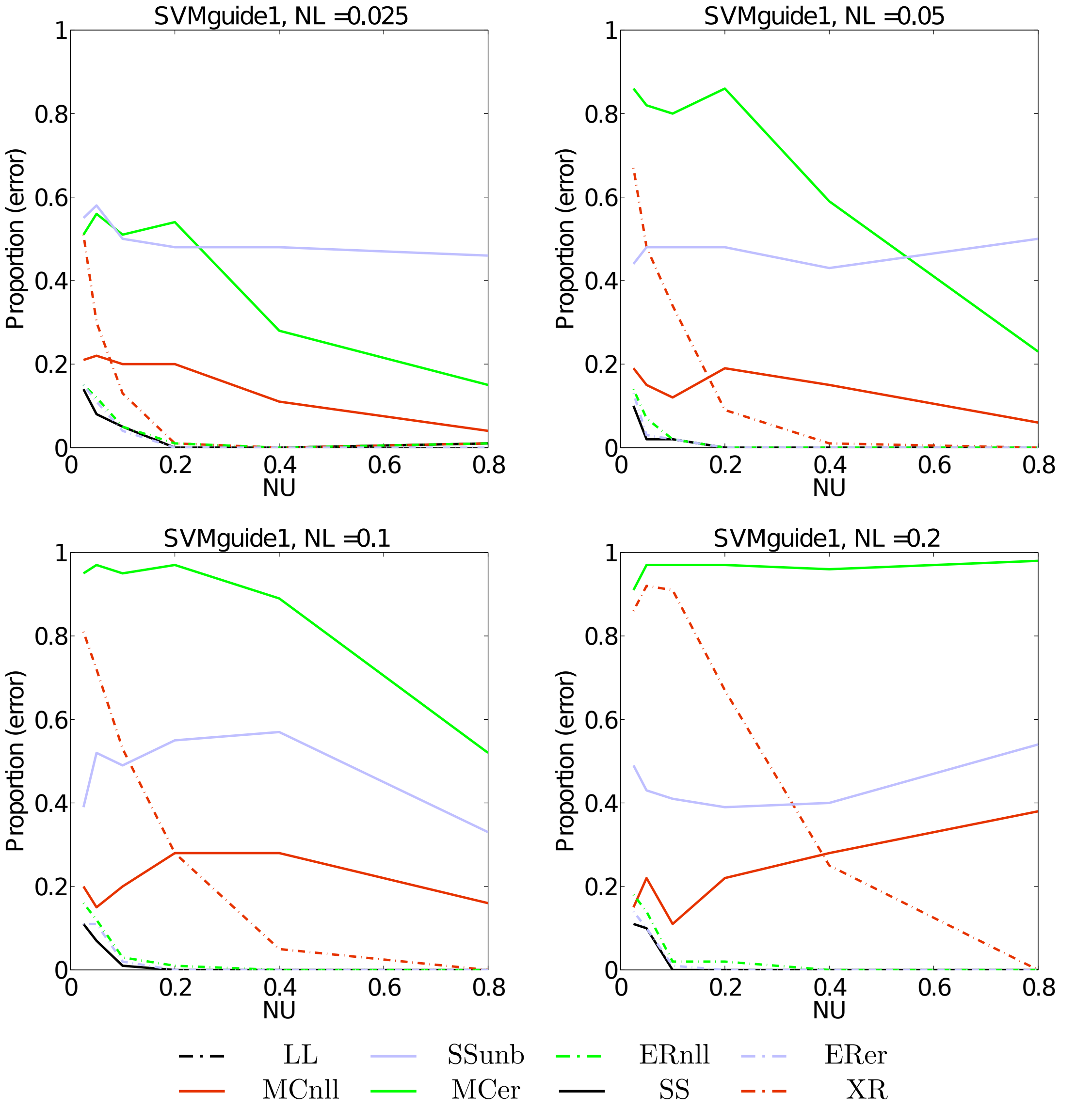

Figure 35: Chance of error rate improvement VS labelled data alone - SVMguide data set

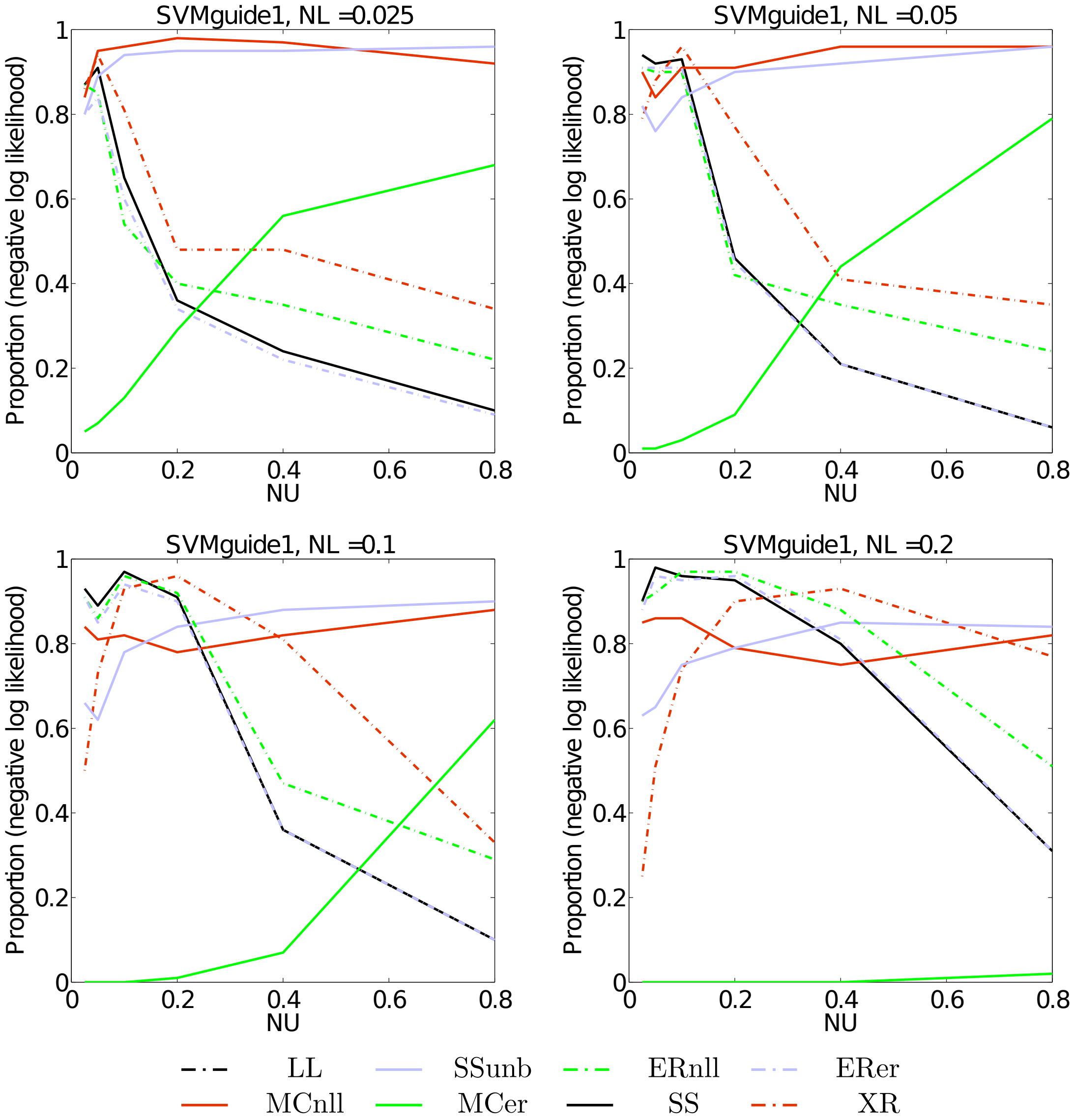

Figure 36: Chance of log likelihood improvement VS labelled data alone - SVMguide data set

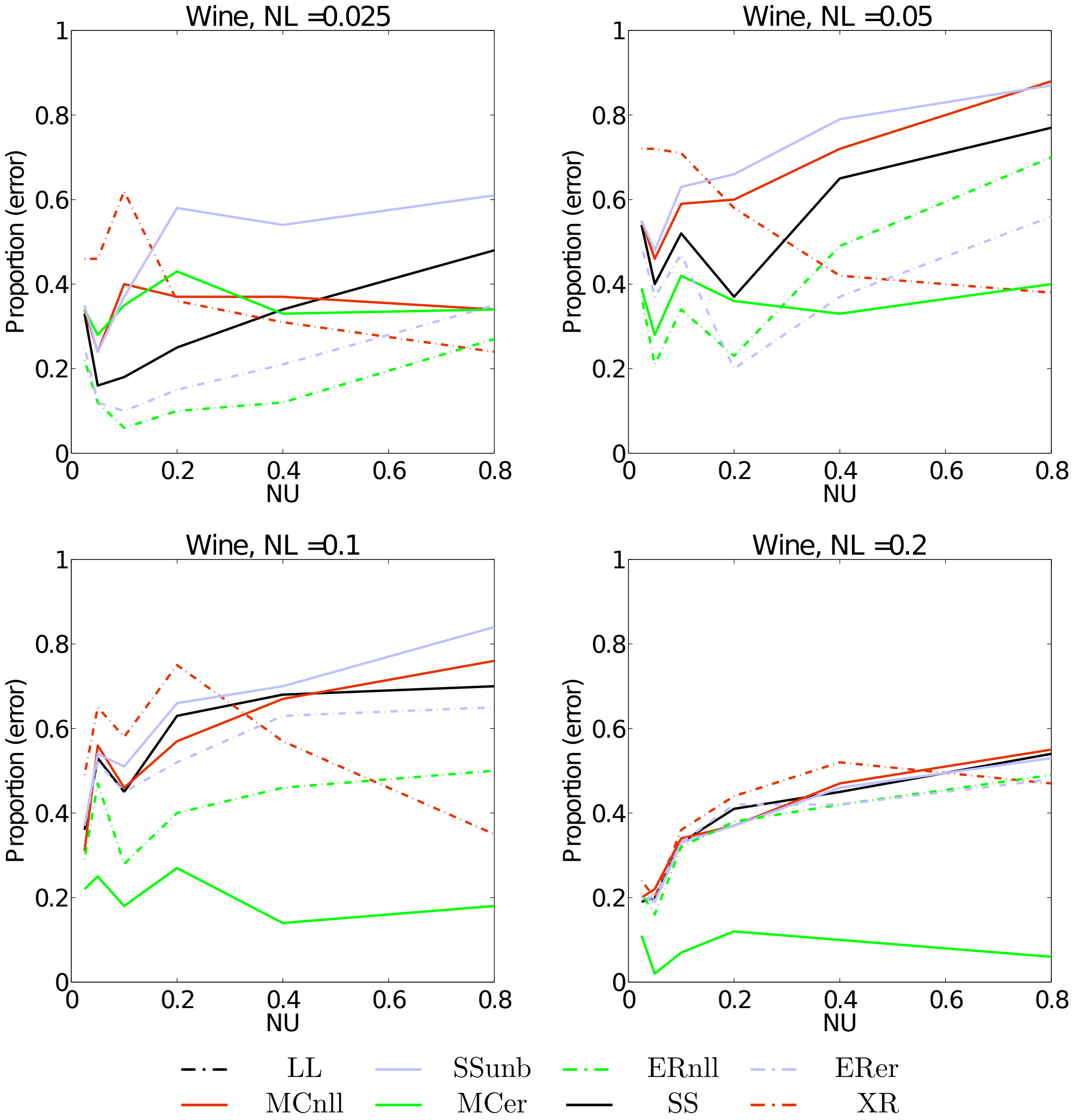

Figure 37: Chance of error rate improvement VS labelled data alone - Wine data set

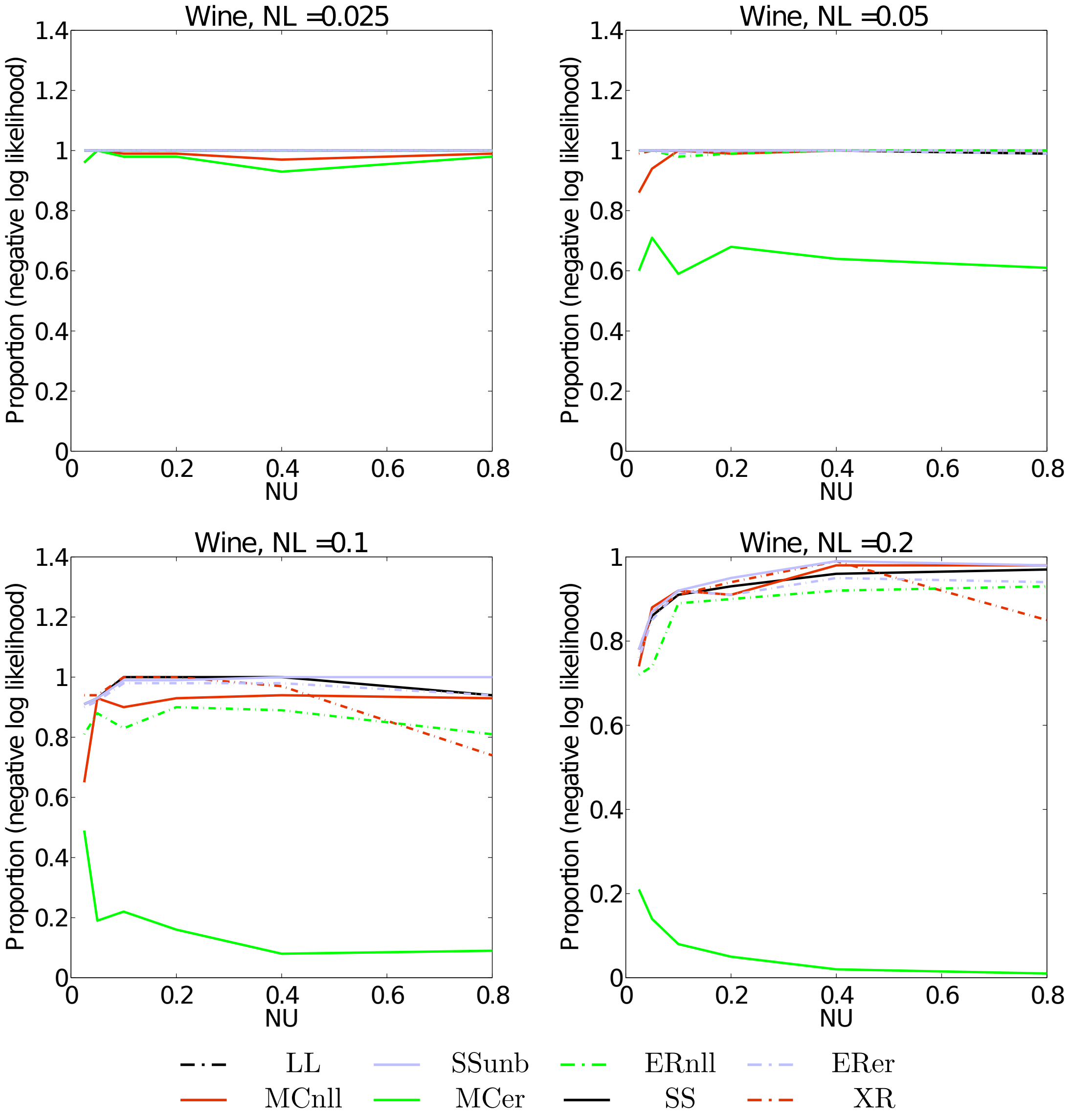

Figure 38: Chance of log likelihood improvement VS labelled data alone - Wine data set

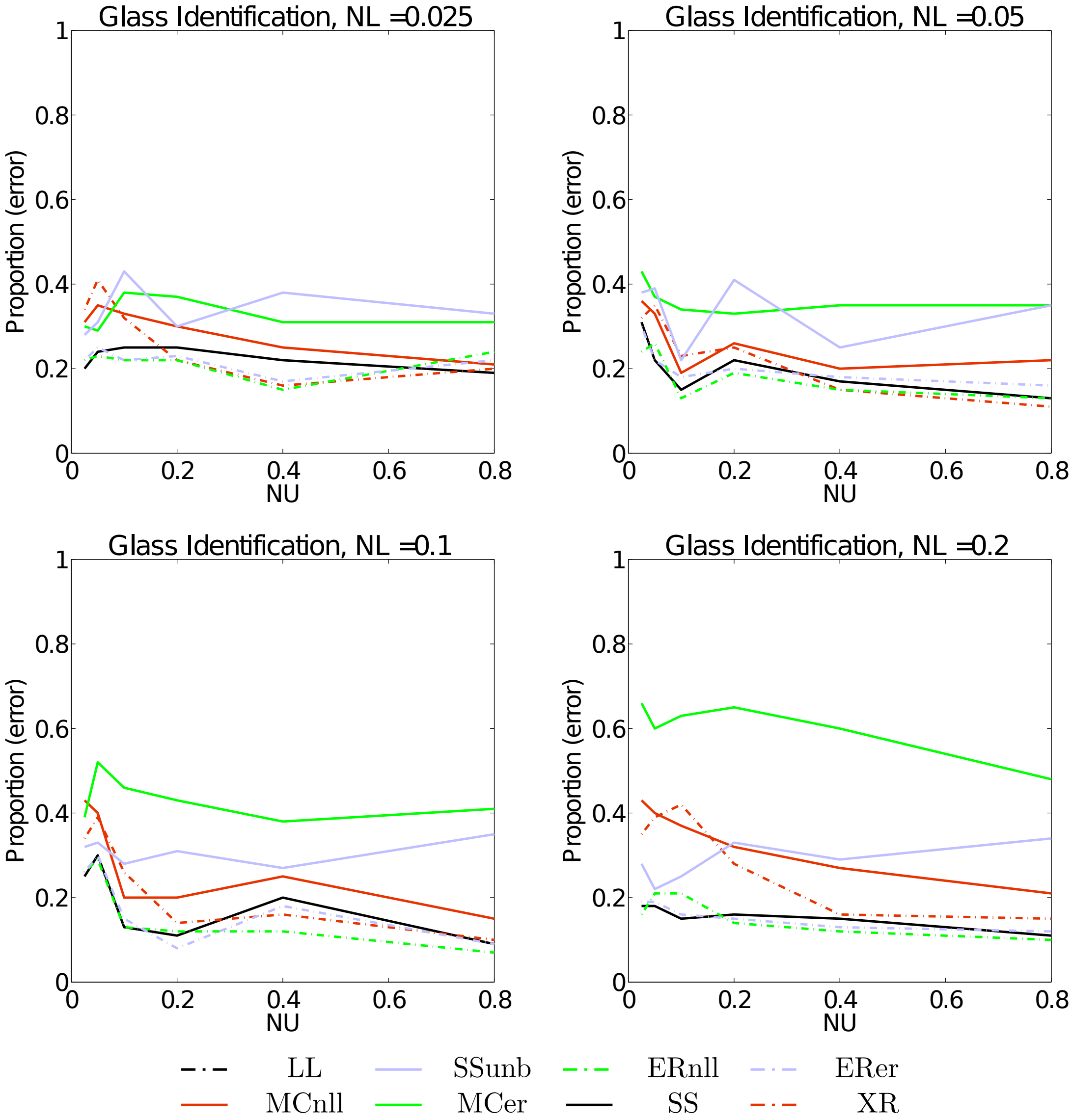

Figure 39: Chance of error rate improvement VS labelled data alone - Glass data set

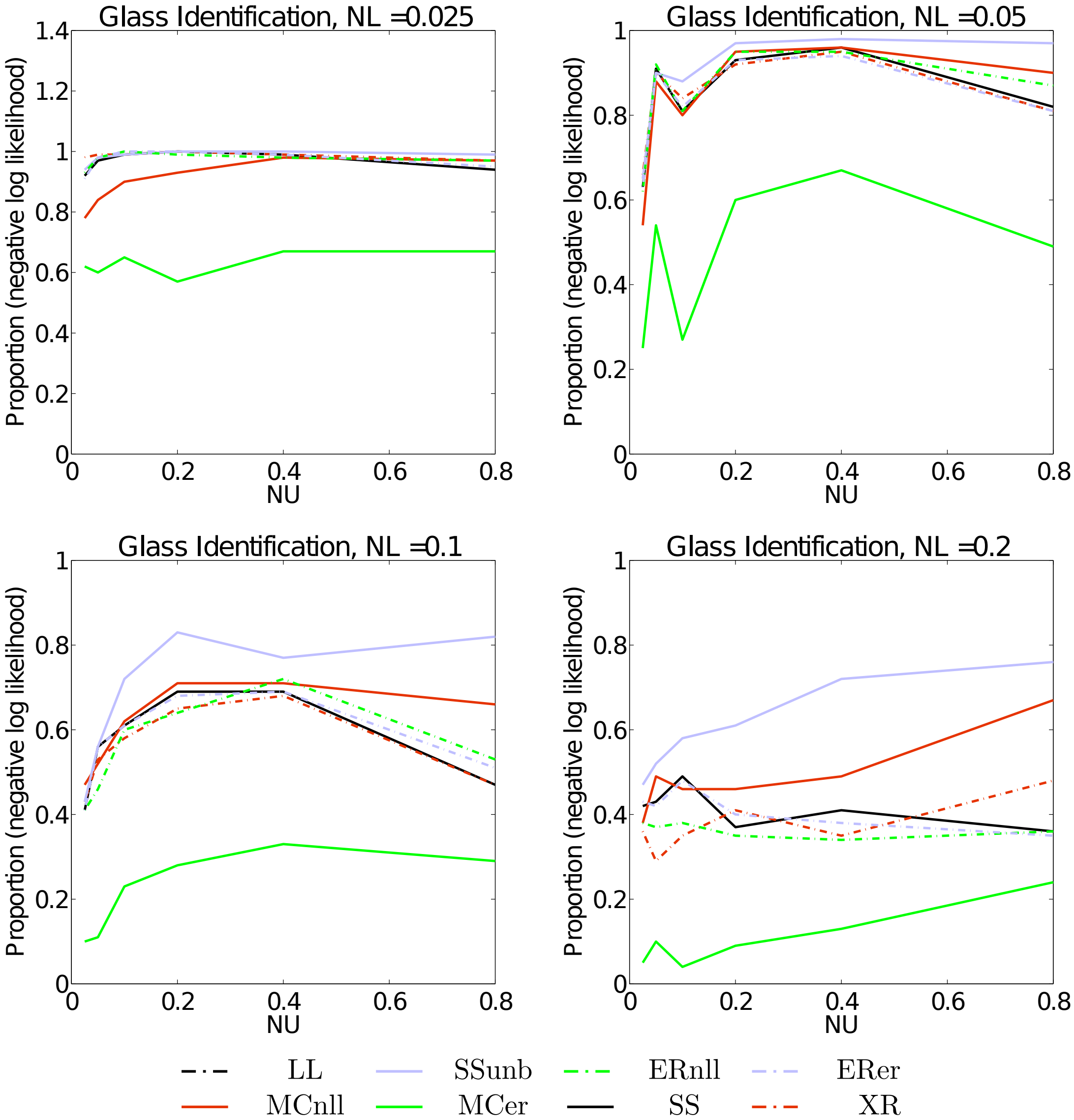

Figure 40: Chance of log likelihood improvement VS labelled data alone - Glass data set

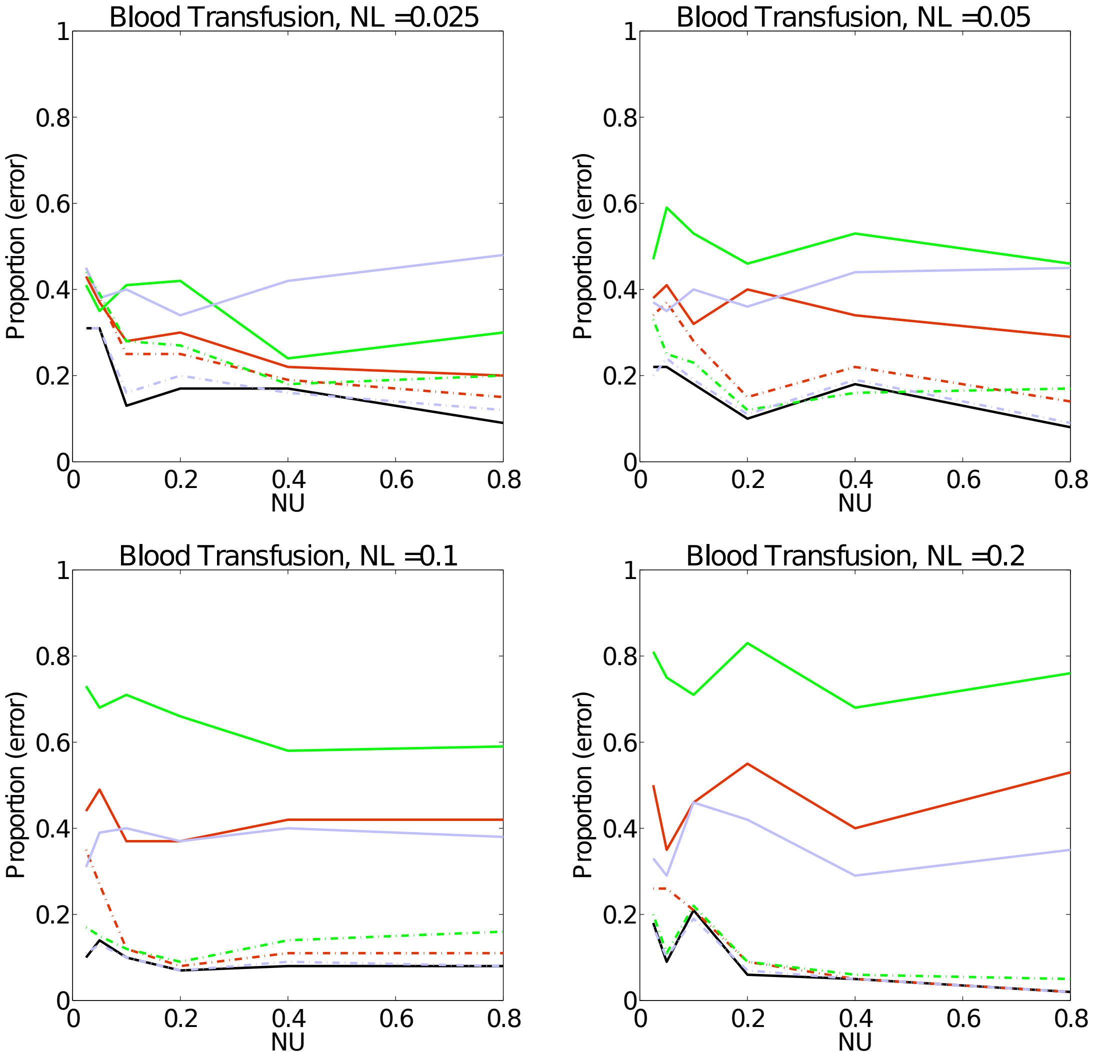

Figure 41: Chance of error rate improvement VS labelled data alone - Blood data set

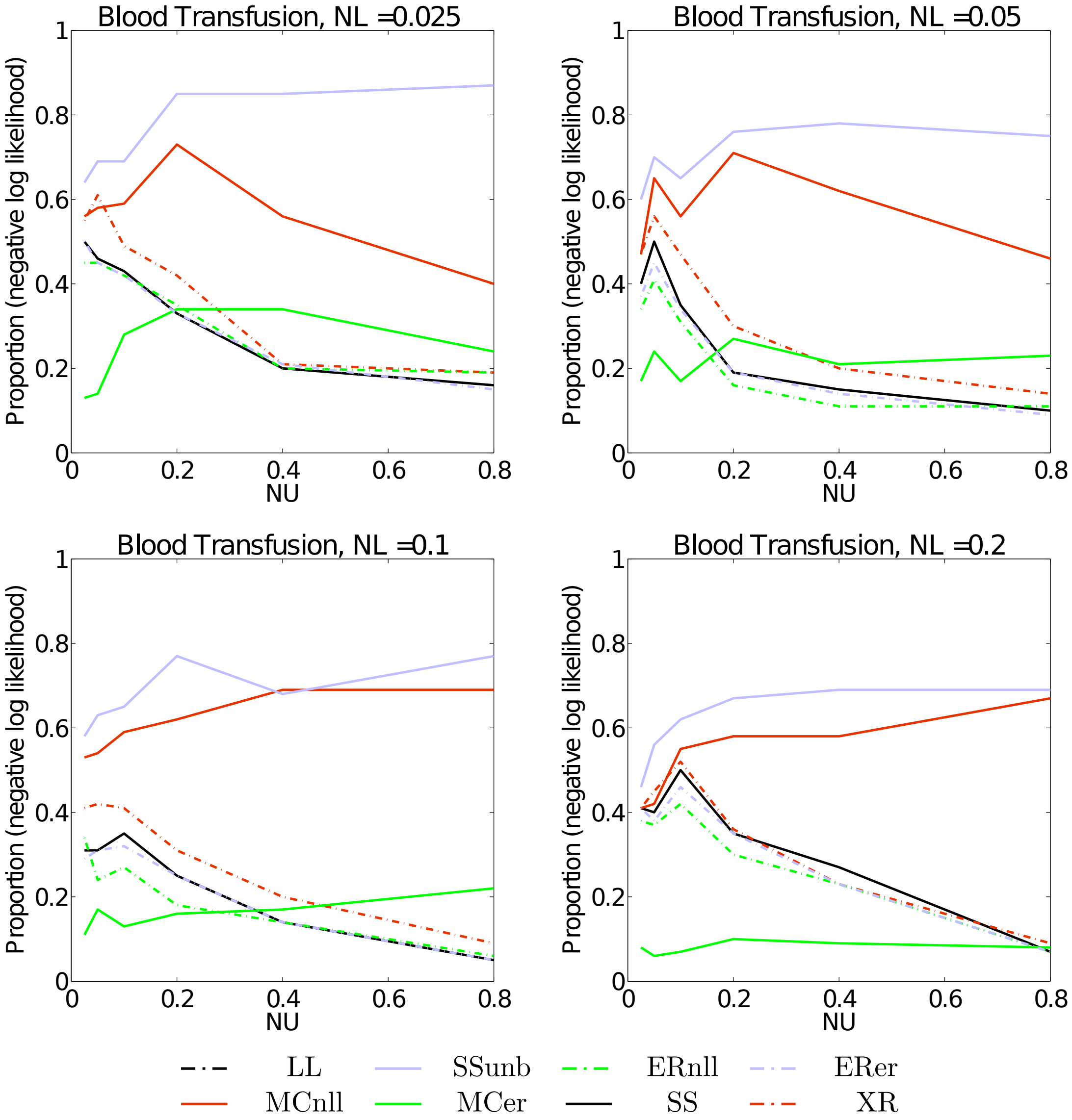

Figure 42: Chance of log likelihood improvement VS labelled data alone - Blood data set

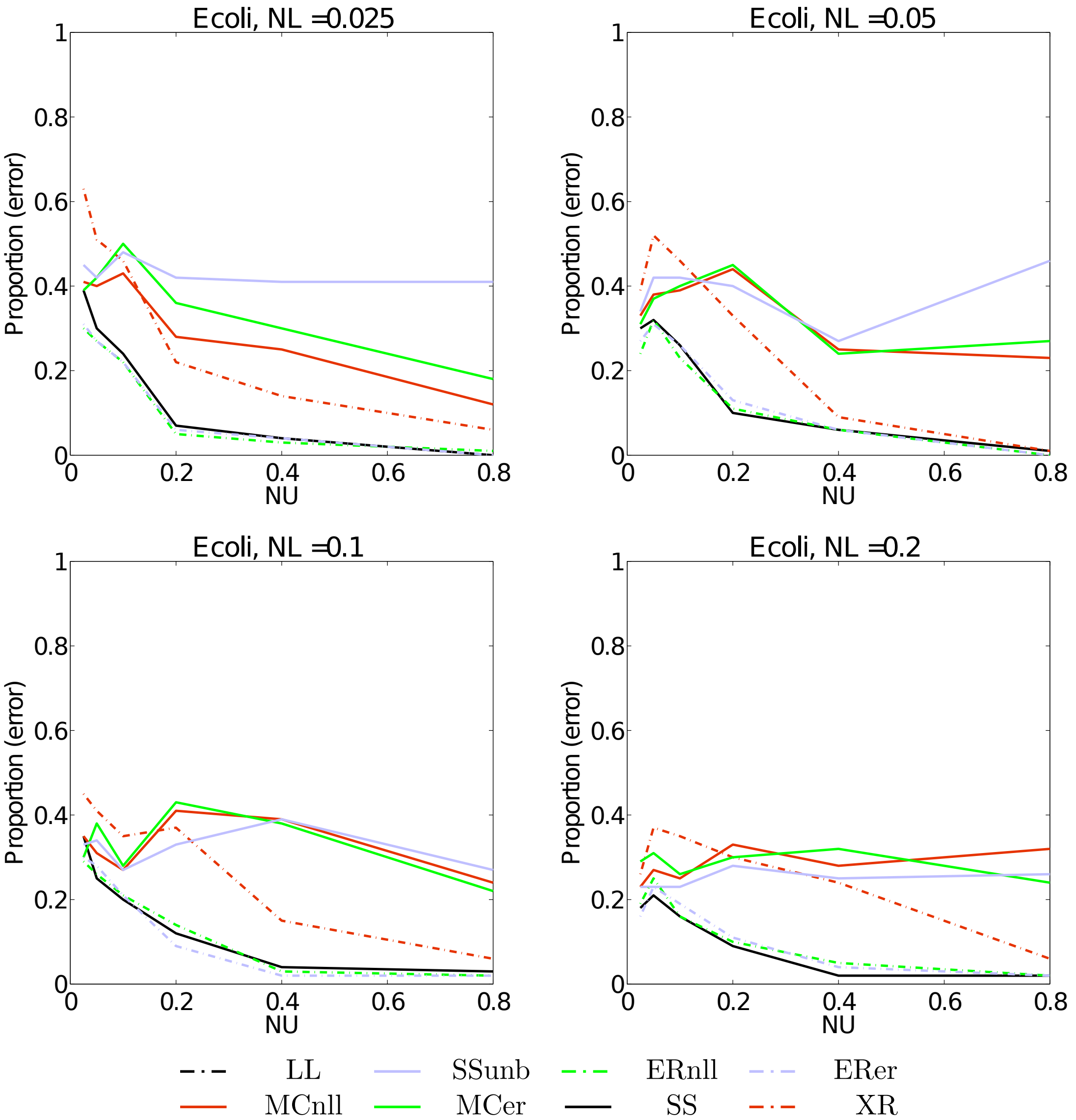

Figure 43: Chance of error rate improvement VS labelled data alone - Ecoli data set

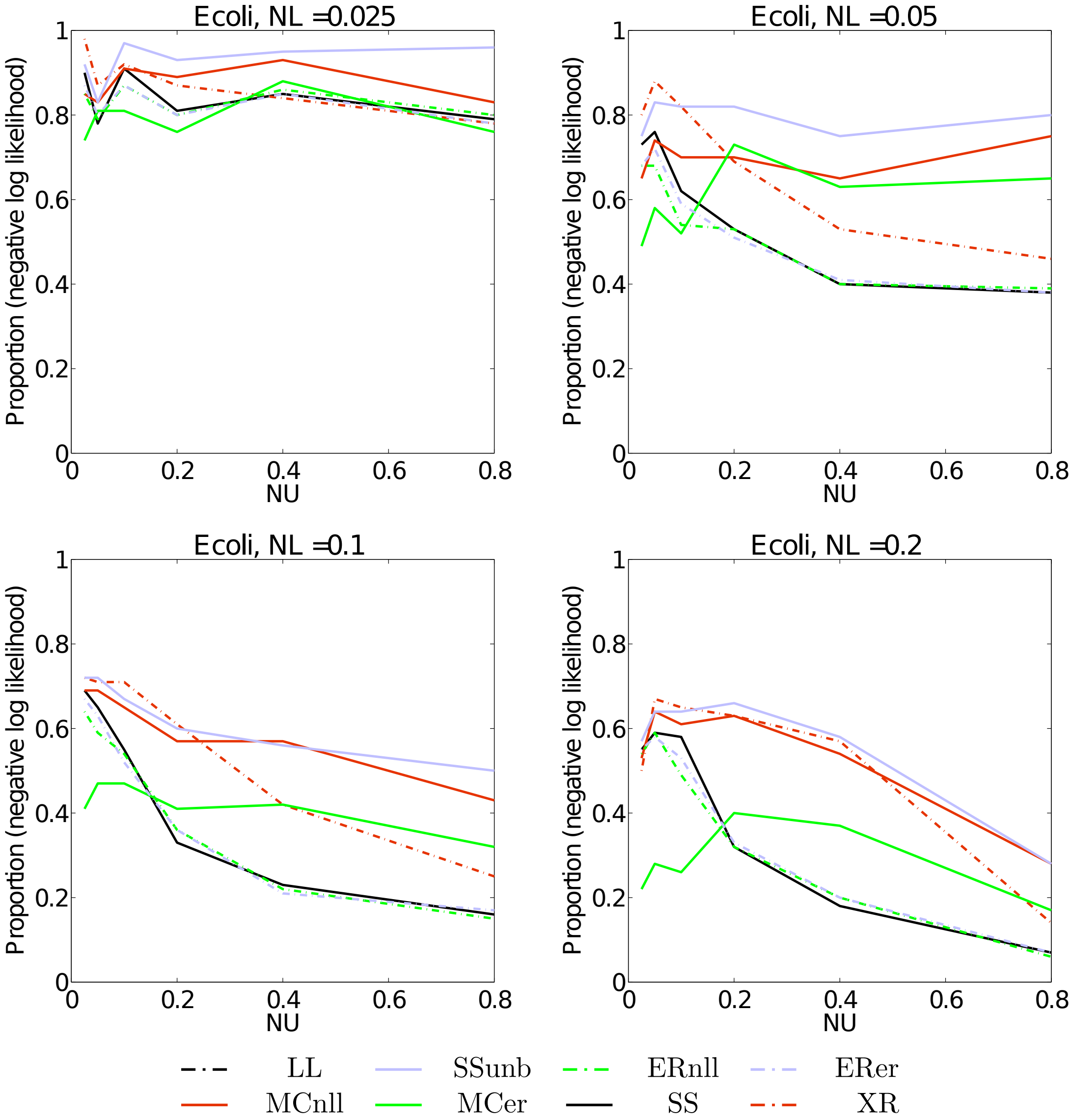

Figure 44: Chance of log likelihood improvement VS labelled data alone - Ecoli data set

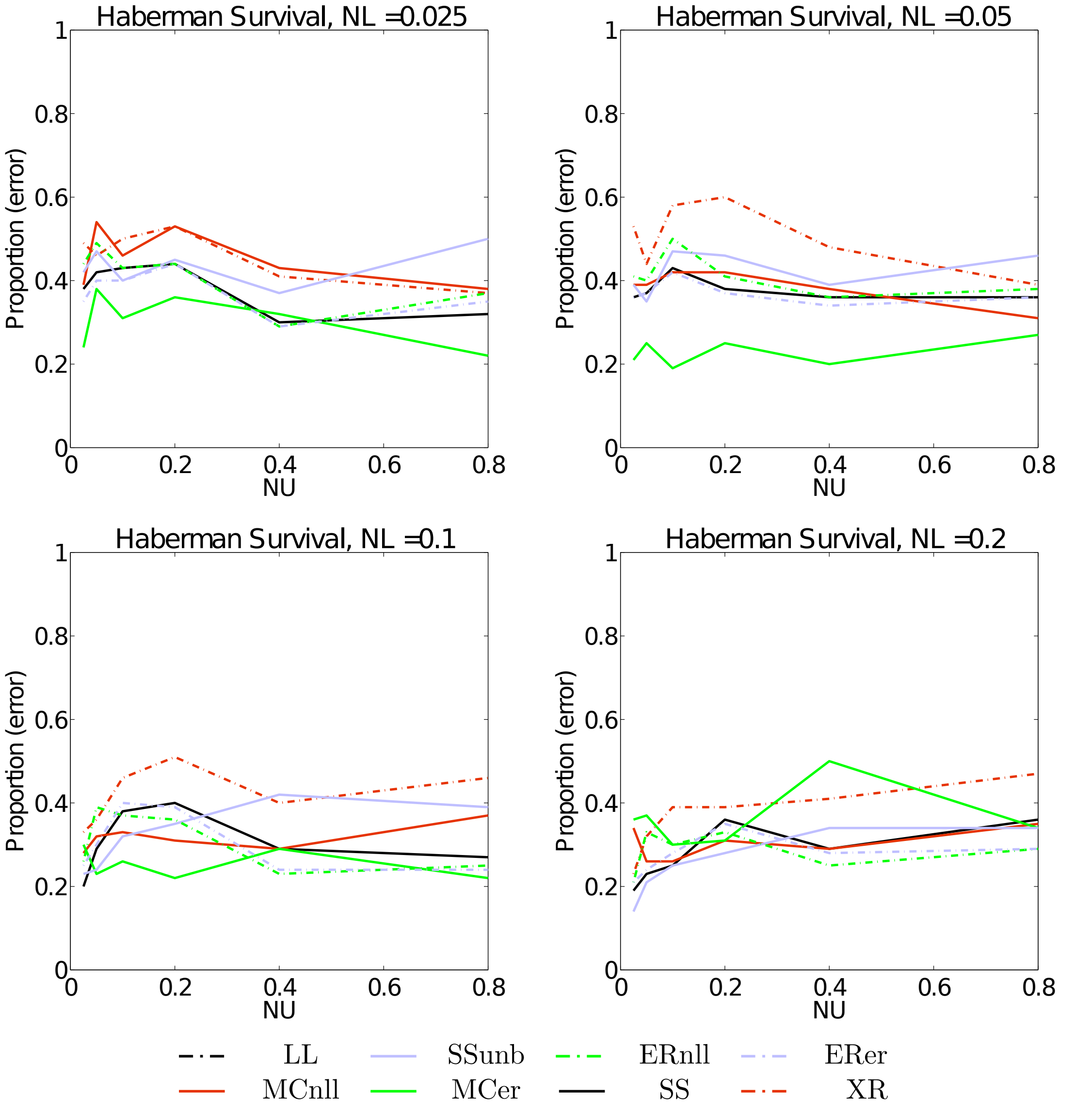

Figure 45: Chance of error rate improvement VS labelled data alone - Haberman data set

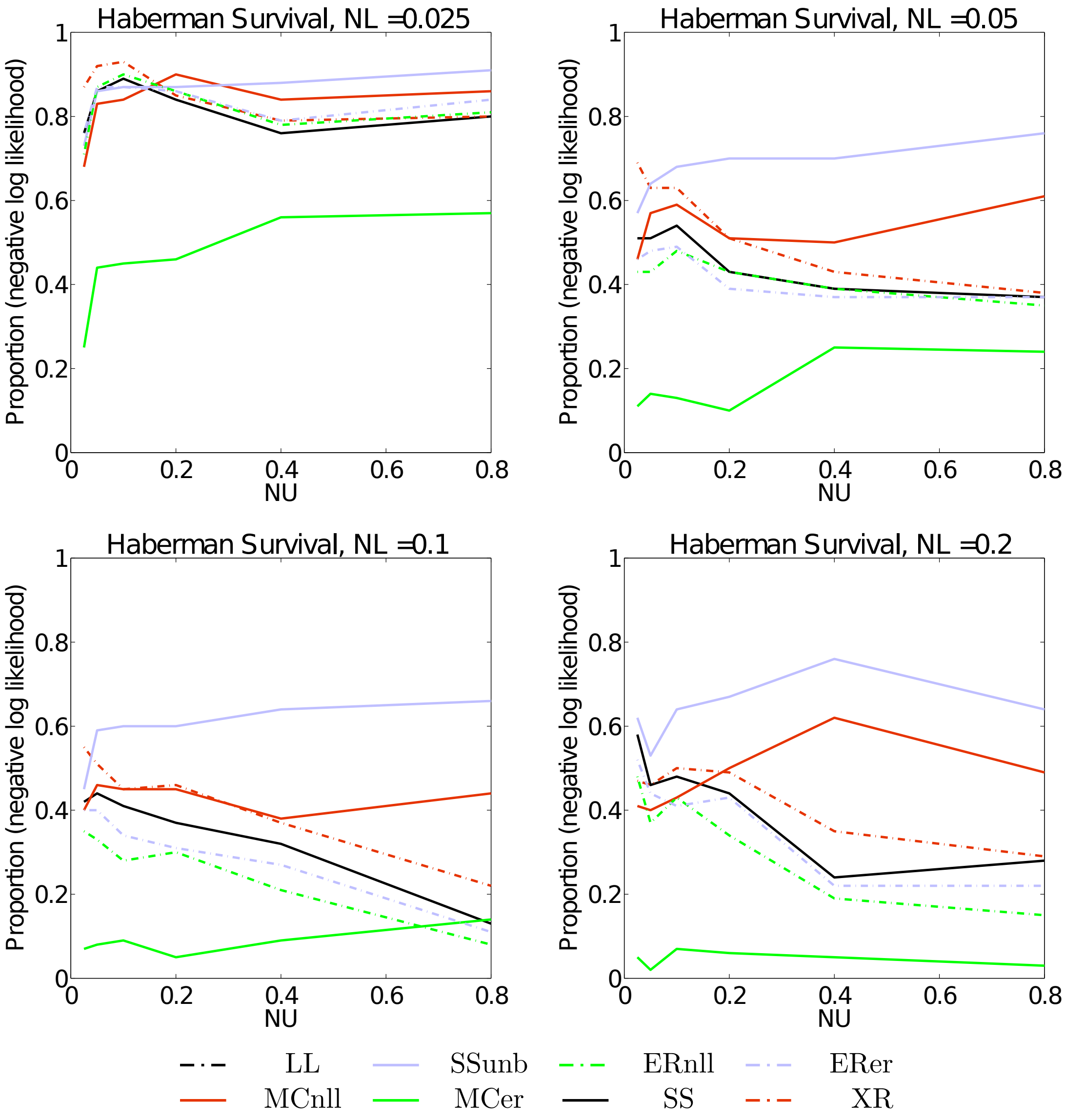

Figure 46: Chance of log likelihood improvement VS labelled data alone - Haberman data set

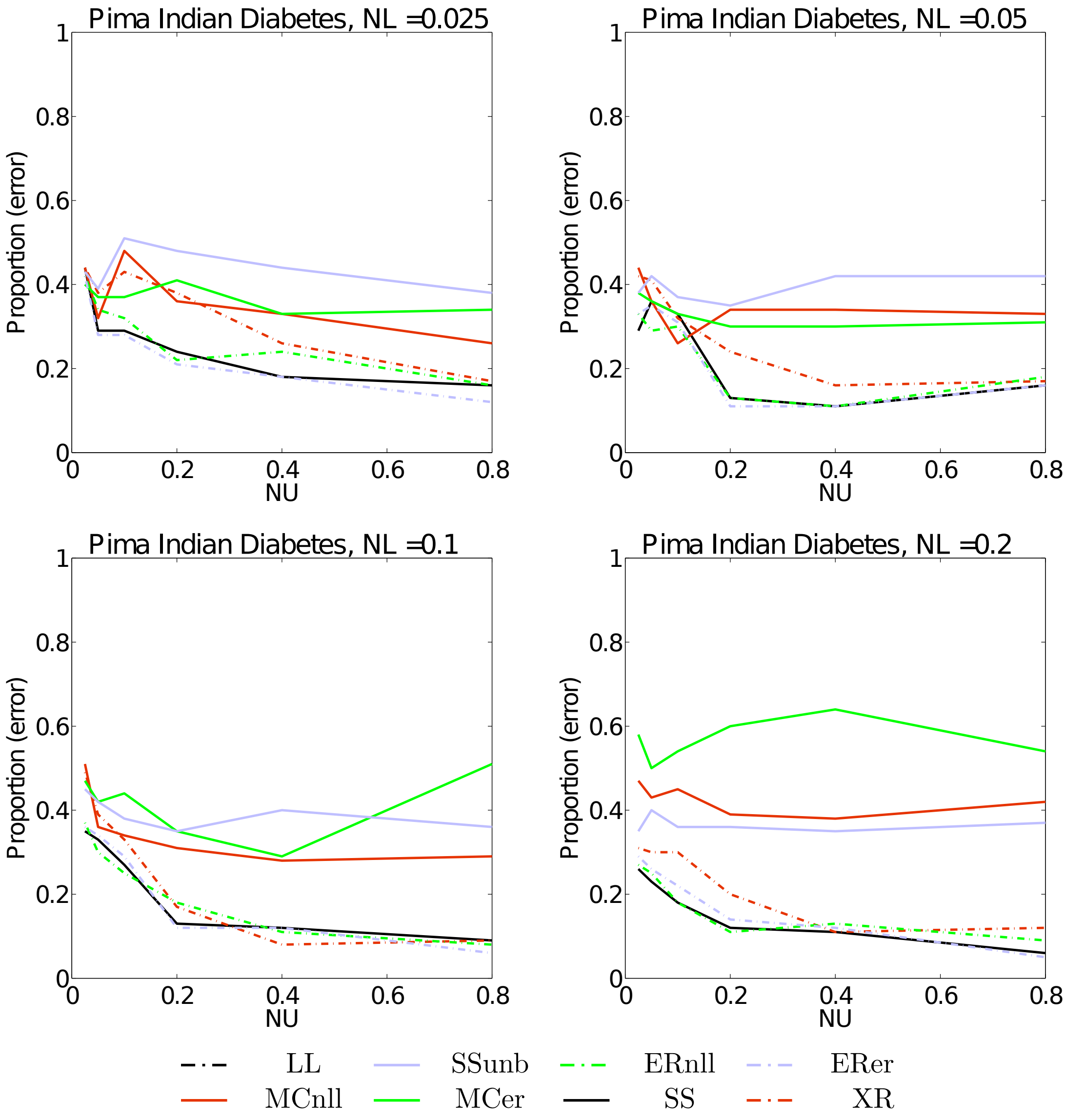

Figure 47: Chance of error rate improvement VS labelled data alone - Pima Indian data set

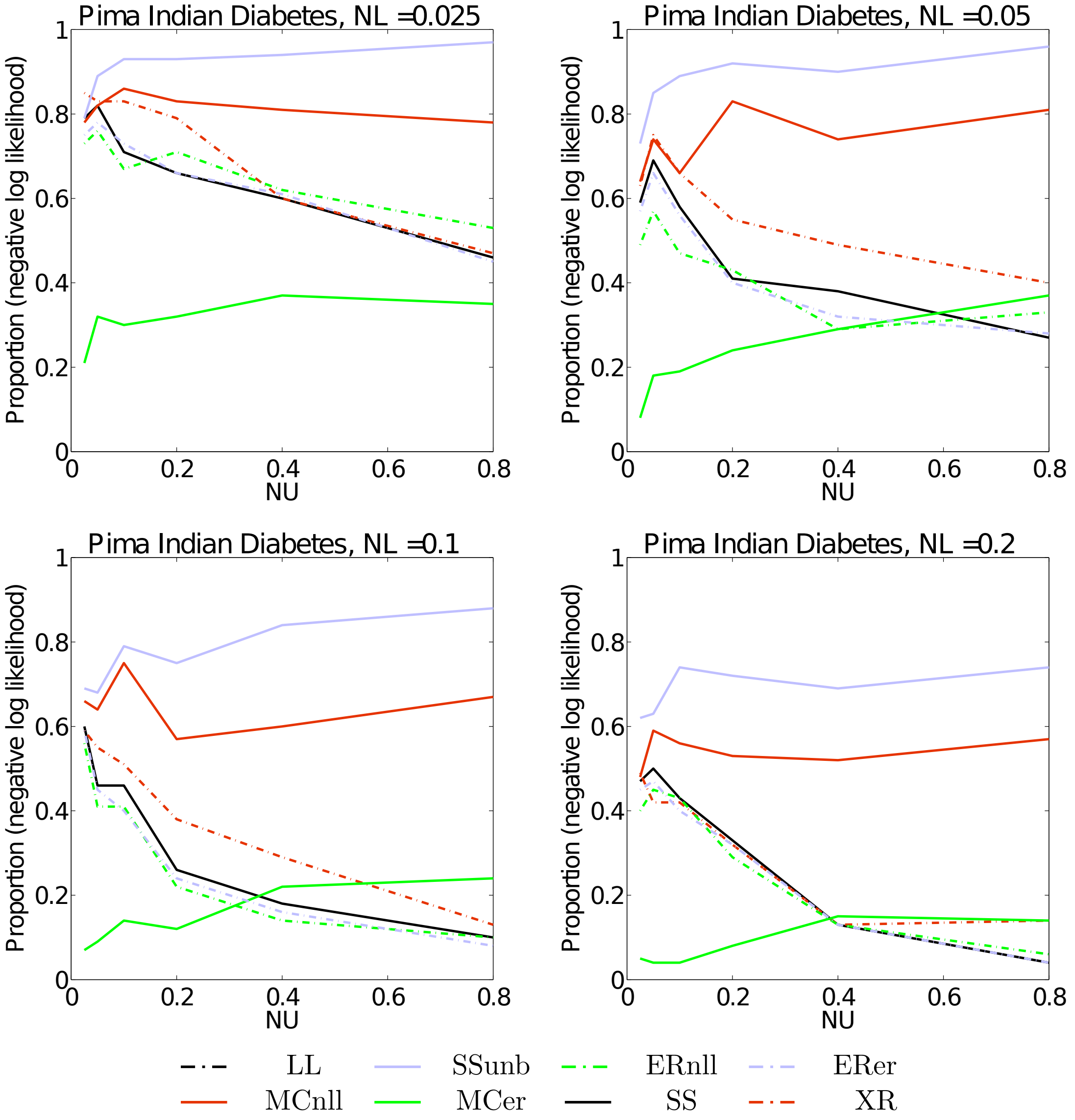

Figure 48: Chance of log likelihood improvement VS labelled data alone - Pima Indian data set

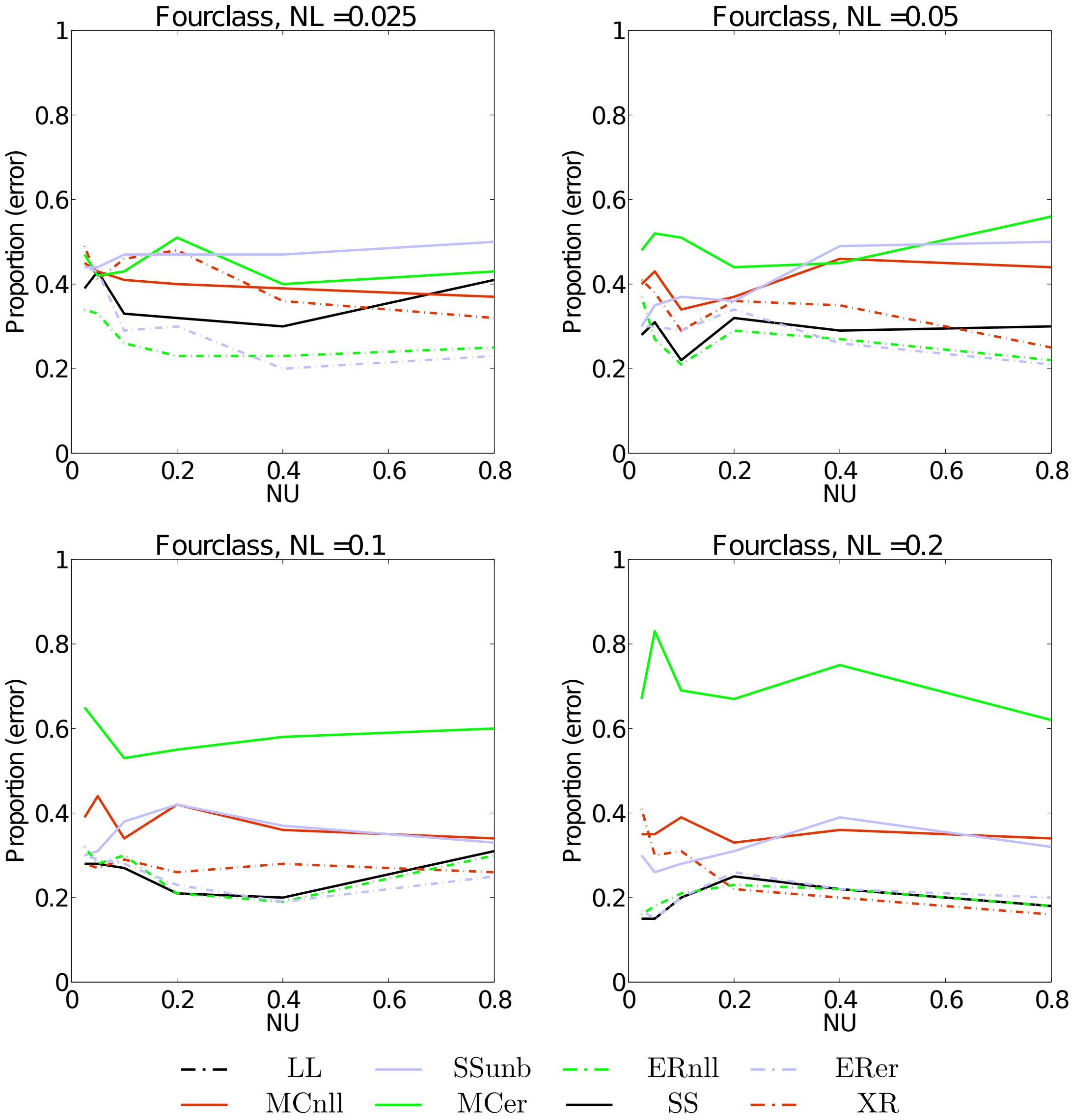

Figure 49: Chance of error rate improvement VS labelled data alone - Fourclass data set

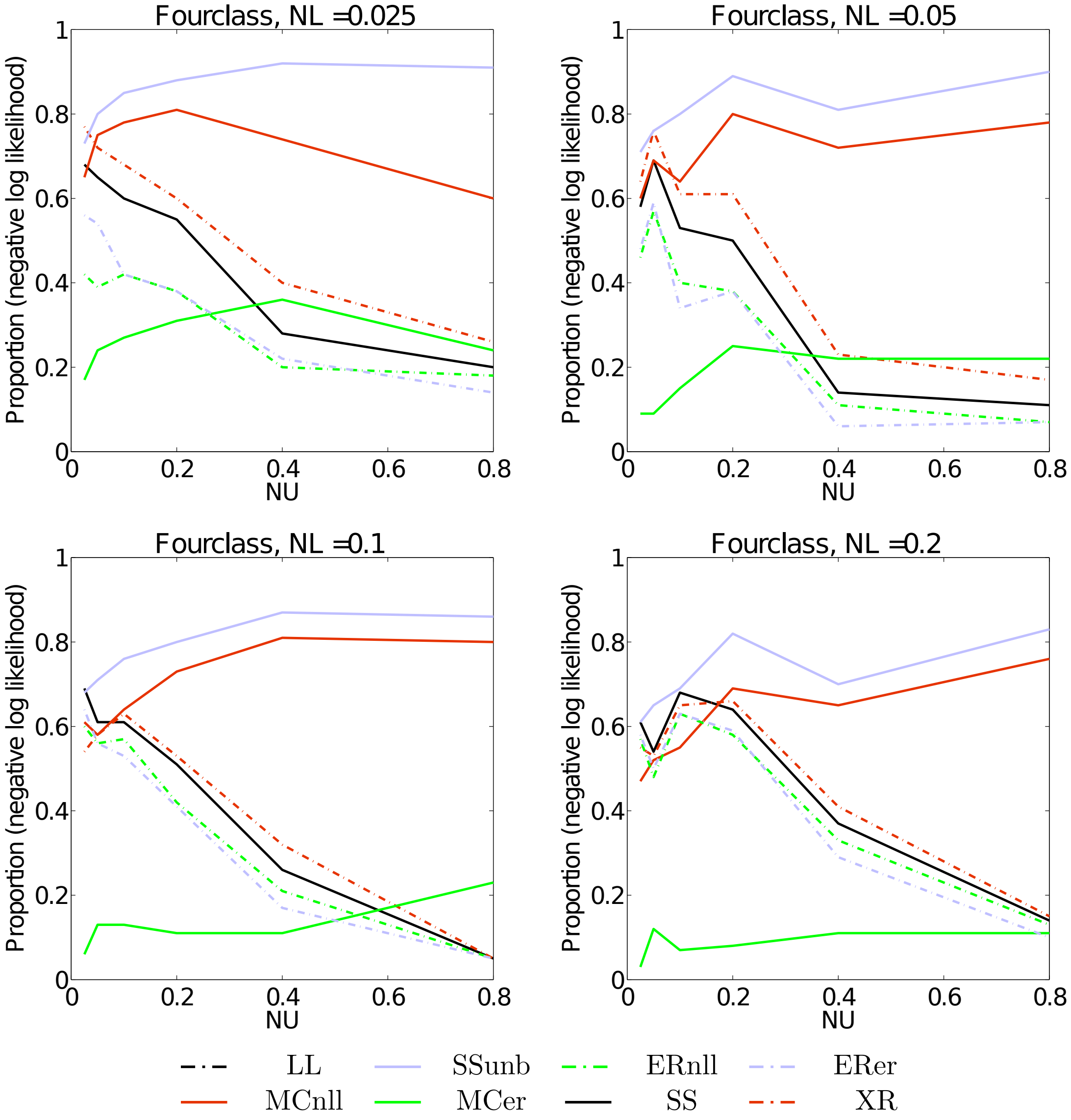

Figure 50: Chance of log likelihood improvement VS labelled data alone - Fourclass data set

References

M. Balcan and A. Blum. An augmented PAC model for semi-supervised learning. In Oliver Chapelle, B. Schölkopf, and A. Zien, editors, *Semi-Supervised Learning*, pages 383–404. MIT Press, 2005.

C. Beecks, A. Ivanescu, S. Kirchhoff, and T. Seidl. Modeling image similarity by gaussian mixture models and the signature quadratic form distance. In *Computer Vision (ICCV), 2011 IEEE International Conference On*, pages 1754–1761, 2011.

C. Bishop. *Pattern Recognition And Machine Learning.* Springer, 2006.

A. Blum and N. Balcan. A discriminative model for semi-supervised learning. In *Journal Of The ACM (JACM)*, volume 57, pages 19:1–19:46, 2010.

A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings Of The Workshop On Computational Learning Theory*, pages 92–100, 1998.

V. Castelli and T. Cover. On the exponential value of labeled samples. *Pattern Recognition Letters*, 16(1):105 – 111, 1995.

V. Castelli and T. Cover. The relative value of labeled and unlabeled samples in pattern recognition with an unknown mixing parameter. *Information Theory, IEEE Transactions on*, 42(6):2102 – 2117, 1996.

F. Cozman and I. Cohen. Unlabeled data can degrade classification performance of generative classifiers. In *Fifteenth International Florida Artificial Intelligence Society Conference*, pages 327–331, 2002.

F. Cozman and I. Cohen. Risks of semi-supervised learning: How unlabelled data can degrade performance of generative classifiers. In O. Chapelle, B. Schölkopf, and A. Zien, editors, *Semi-Supervised Learning*, pages 57–72. MIT press, 2006.

F. Cozman, I. Cohen, M. Cirelo, and E. Politécnica. Semi-supervised learning of mixture models. In *Proceedings Of The 20th International Conference On Machine Learning (ICML)*, pages 99–106, 2003.

J. Dillon, K. Balasubramanian, and G. Lebanon. Asymptotic analysis of generative semi-supervised learning. In *Proceedings Of The 27th International Conference On Machine Learning (ICML)*, 2010.

G. Druck, C. Pal, A. McCallum, and X. Zhu. Semi-supervised classification with hybrid generative/discriminative methods. In *KDD '07: Proceedings Of The 13th ACM SIGKDD International Conference On Knowledge Discovery And Data Mining*, pages 280–289, 2007.

A. Frank and A. Asuncion. UCI machine learning repository, 2010. URL `http://archive.ics.uci.edu/ml`.

Y. Grandvalet and Y. Bengio. Entropy Regularization. In *Semi-Supervised Learning*, pages 151–168. MIT Press, 2006.

T. Ho and E. Kleinberg. Building projectable classifiers of arbitrary complexity. In *International Conference On Pattern Recognition (ICPR)*, volume 2, pages 880–885, 1996.

C. Hsu, C. Chang, and C. Lin. A practical guide to support vector classification. Technical report, Department of Computer Science, National Taiwan University, 2003.

G. Hughes. On the mean accuracy of statistical pattern recognizers. *Information Theory, IEEE Transactions On*, 14(1):55 – 63, 1968.

T. Jaakkola and D. Haussler. Exploiting generative models in discriminative classifiers. In *Proceedings Of The 1998 Conference On Advances In Neural Information Processing Systems II*, pages 487–493, 1999.

E. Jaynes. *Probability Theory.* Cambridge University Press, 2003.

H. Kang, S. Yoo, and D. Han. Senti-lexicon and improved naïve bayes algorithms for sentiment analysis of restaurant reviews. *Expert Systems With Applications*, 39(5):6000 – 6010, 2012.

B. Krishnapuram, D. Williams, Y. Xue, A. Hartemink, L. Carin, and M. Figueiredo. On semi-supervised classification. In *Advances In Neural Information Processing Systems (NIPS)*, pages 721–728, 2005.

J. Lagarias, J. Reeds, M. Wright, and P. Wright. Convergence properties of the Nelder-Mead simplex method in low dimensions. *SIAM Journal Of Optimization*, 9(1):112–147, 1998.

J. Lasserre, C. Bishop, and T. Minka. Principled hybrids of generative and discriminative models. In *Computer Vision And Pattern Recognition (CVPR)*, volume 1, pages 87 – 94, 2006.

J. Lücke and J. Eggert. Expectation truncation and the benefits of preselection in training generative models. In *Journal Of Machine Learning Research (JMLR)*, pages 2855–2900, October 2010.

D. MacKay. *Information Theory, Inference, And Learning Algorithms.* Cambridge University Press, 2003.

G. Mann and A. McCallum. Simple, robust, scalable semi-supervised learning via Expectation Regularization. In *Proceedings Of The 24th International Conference On Machine Learning (ICML)*, pages 593–600, 2007.

A. McCallum, C. Pal, G. Druck, and X. Wang. Multi-conditional learning: Generative/discriminative training for clustering and classification. In *National Conference On Artificial Intelligence*, pages 433–439, 2006.

A. Ng and M. Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and Naive Bayes. *Advances In Neural Information Processing Systems (NIPS)*, 2:841–848, 2002.

K. Nigam, A. McCallum, S. Thrun, and T. Mitchell. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39:103–134, 2000.

J. Nocedal and S. Wright. *Numerical Optimization, Springer Series In Operations Research.* Springer-Verlag, 1999.

J. Ratsaby and S. Venkatesh. Learning from a mixture of labeled and unlabeled examples with parametric side information. In *Conference On Learning Theory (COLT)*, pages 412–417, 1995.

I. Rauschert and R. Collins. A generative model for simultaneous estimation of human body shape and pixel-level segmentation. In *Proceedings Of The 12th European Conference On Computer Vision - Volume Part V*, European Conference On Computer Vision (ECCV), pages 704–717. Springer-Verlag, 2012.

S. Rosset, J. Zhu, H. Zou, and T. Hastie. A method for inferring label sampling mechanisms in semi-supervised learning. In *Advances In Neural Information Processing Systems (NIPS)*, volume 17, pages 1161 – 1168, 2005.

B. Shahshahani and D. Landgrebe. The effect of unlabeled samples in reducing the small sample size problem and mitigating the Hughes phenomenon. *Geoscience And Remote Sensing, IEEE Transactions on*, 32(5):1087 –1095, 1994.

A. Subramanya and J. Bilmes. Soft-supervised learning for text classification. In *Proceedings Of The Conference On Empirical Methods In Natural Language Processing*, pages 1090–1099, 2008.

A. Subramanya and J. Bilmes. Entropic graph regularization in non-parametric semi-supervised classification. In *Advances In Neural Information Processing Systems (NIPS)*, December 2009.

U. Syed and B. Taskar. Semi-supervised learning with adversarially missing label information. In *Advances In Neural Information Processing Systems (NIPS)*, pages 2244–2252, 2010.

M. Szummer and T. Jaakkola. Information Regularization with partially labeled data. In *Advances In Neural Information Processing Systems (NIPS)*, 2002.

V. Vapnik. *Statistical Learning Theory.* John Wiley & Sons, Inc, 1998.

J. Wang, X. Shen, and W. Pan. On transductive support vector machines. In *Prediction And Discovery.* American Mathematical Society, 2007.

T. Yang and C. Priebe. The effect of model misspecification on semi-supervised classification. *Pattern Analysis And Machine Intelligence (PAMI)*, 33:2093–2103, 2011.

I. Yeh, K. Yang, and T. Ting. Knowledge discovery on RFM model using Bernoulli sequence. *Expert Systems With Applications*, 36(3, Part 2):5866 – 5871, 2009.

T. Zhang. The value of unlabeled data for classification problems. In *International Conference On Machine Learning (ICML)*, pages 1191–1198, 2000.

X. Zhu. Semi-supervised learning literature survey. Technical report, Department of Computer Sciences, University of Wisconsin, Madison, 2005.

F. Zhuang, P. Luo, Z. Shen, Q. He, Y. Xiong, Z. Shi, and H. Xiong. Mining distinction and commonality across multiple domains using generative model for text classification. *Knowledge And Data Engineering, IEEE Transactions On*, 24(11):2025–2039, 2012.