Multi-class Semi-supervised Learning With The ϵ-truncated Multinomial Probit Gaussian Process

Simon Rogers SROGERS@DCS.GLA.AC.UK
Mark Girolami GIROLAMI@DCS.GLA.AC.UK
Department of Computing Science
University of Glasgow
G12 8QQ, UK

Editor: Neil D. Lawrence, Anton Schwaighofer and Joaquin Quiñonero Candela

Abstract

Recently, the null category noise model has been proposed as a simple and elegant solution to the problem of incorporating unlabeled data into a Gaussian process (GP) classification model. In this paper, we show how this binary likelihood model can be generalised to the multi-class setting through the use of the multinomial probit GP classifier. We present a Gibbs sampling scheme for sampling the GP parameters and also derive a more efficient variational updating scheme. We find that the performance improvement is roughly consistent with that observed in binary classification and that there is no significant difference in classification performance between the Gibbs sampling and variational schemes.

1. Introduction

In machine learning, we are often faced with the problem of classification — devising a mapping between an input vector $\mathbf{x}$ and some label t that can either be binary ($t \in \{-1, 1\}$) or discrete over some finite set of classes ($t \in \{1, \ldots, \mathcal{K}\}$). Generally, this mapping is inferred from some set of *training* vectors $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^T$ and their associated labels $\mathbf{t} = [t_1, \ldots, t_N]^T$. Recently, there has been considerable research interest in whether the predictions made by such a model could be improved via the inclusion of additional training vectors for which the true label is unknown. Such a technique would be useful in many fields where expert labeling of training examples is costly, for example prediction of protein function from structural information and classification of documents and images on the internet.

Classification algorithms based on Gaussian Processes (GPs) (e.g. Williams and Barber, 1998) are becoming increasingly popular due to the computational and performance benefits of their non-parametric nature and the flexibility provided through the large number of covariance functions available. However, like all discriminative probabilistic classifiers, there is an inherent problem with the inclusion of unlabeled data. Specifically, through direct modeling of $p(t_n|\mathbf{x}_n)$ (as is the case with a discriminative model) with no prior assumptions on the density of the input data, the unlabeled data will have no influence on the inferred decision function. This point is illustrated nicely via a graphical representation in Lawrence and Jordan (2006) and an interesting discussion on this subject (outside the realm of GPs) is given in Lasserre et al. (2006). Incorporating unlabeled data in a generative framework is

far more straightforward as unlabeled points can easily be used to improve a model of the density, $p(\mathbf{x})$ from which the data is assumed to be sampled. However, in complicated and high dimensional domains, modeling $p(\mathbf{x})$ can be very costly, motivating the development of discriminative classifiers.

Some methods have been proposed to incorporate unlabeled data into GP classifiers (we are interested primarily in GPs here but solutions have been proposed for other discriminative classification frameworks too). For example, in Seeger (2001), a kernel is introduced based on a model of $p(\mathbf{x})$ and Chapelle et al. (2003) propose a solution based on kernels that attempt to enforce the cluster assumption — i.e., decision boundaries should lie in regions of low data density. Additionally, Zhu et al. (2003) show that a semi-supervised method based on Gaussian random fields can be viewed within the framework of Gaussian processes.

In Lawrence and Jordan (2006), the authors take a rather different approach. They show that for binary classification, the problem can be overcome through the use of the *null category noise model* (NCNM). This introduces an extra target category $t_n = 0$ and an additional set of parameters z_n where $z_n = 0$ if the label of $\mathbf{x}_n$ is observed and $z_n = 1$ otherwise. Crucially, the model is also constrained such that $p(z_n = 1|t_n = 0) = 0$ i.e., there can be no un-labelled points in the null category. This has the effect of creating a region in the input space inside which no points (labeled or unlabeled) exist. This forces the decision boundary to exist in regions of low data density — i.e., it implicitly enforces the clustering assumption. It is obvious that under these conditions, unlabeled data can indeed influence the position of the decision boundary and hence future predictions. There are obvious similarities between this null region and the *margin* in the context of support vector machines.

In this paper, we will show how the null-category noise model is closely related to the binary variant of a recently published GP classification algorithm based on the multinomial probit likelihood function (Girolami and Rogers, 2006). With this in mind, we will proceed to show how combining the idea of a null category with the multi-class probit GP results in a straight-forward algorithm for discriminative semi-supervised learning over multiple classes.

The remainder of the paper is organised as follows. In section 2 we will review the null category noise model used by Lawrence and Jordan (2006). In section 3 we will show how this model is closely related to the probit GP and hence show how the idea can be extended to perform multi-class classification through Markov chain Monte-Carlo. An efficient variational approximation is provided in section 4, in section 5 we provide some experimental results and in section 6 draw some conclusions.

2. The null-category noise model

The standard (binary) GP classification model can be defined as follows

$$p(\mathbf{t}|\mathbf{X}) = \int p(\mathbf{t}|\mathbf{m})p(\mathbf{m}|\mathbf{X})d\mathbf{m}$$

where $\mathbf{m}|\mathbf{X} \sim GP(\mathbf{0}, \mathbf{C})$ (i.e. a Gaussian Process with a zero mean and covariance matrix $\mathbf{C}$, the ijth element of which is some covariance function evaluated at $\mathbf{x}_i, \mathbf{x}_j$) and $p(t_n|m_n)$

is the noise model or likelihood, the definition of which distinguishes most GP classification methods. Here we are interested in the null-category noise model which, with the addition of a set of auxiliary variables y_n such that $y_n|m_n \sim \mathcal{N}(m_n, 1)$, is defined in Lawrence and Jordan (2006) as

$$p(t_n|y_n) = \begin{cases} \delta(y_n < -\frac{1}{2}) & \text{for} \quad t_n = -1 \\ \delta(y_n > -\frac{1}{2}) - \delta(y_n > \frac{1}{2}) & \text{for} \quad t_n = 0 \\ \delta(y_n > \frac{1}{2}) & \text{for} \quad t_n = 1 \end{cases}$$

where $\delta(expr) = 1$ if the expression $expr$ is true and zero otherwise.

The introduction of an additional set of fully observed binary indicator variables z_n such that $z_n = 1$ if the label for the nth data point is not observed and $z_n = 0$ otherwise enables us to draw samples from the posterior $p(\mathbf{y}, \mathbf{m}|\mathbf{X}, \mathbf{t}, \mathbf{z})$ using a Gibbs sampler as follows (conditioning on any parameters of the covariance function is implicit). Firstly, standard results for the product of two exponential forms give $\mathbf{m} \sim \mathcal{N}(\Sigma\mathbf{y}, \Sigma)$ where $\Sigma = \mathbf{C}(\mathbf{C}+\mathbf{I}_N)^{-1}$, $\mathbf{C}$ is the GP covariance and $\mathbf{I}_N$ is an $N \times N$ identity matrix. More interesting however is the sampling distribution for each y_n. These can be split into two cases, those where the label is observed ($z_n = 0$) and those where it isn't ($z_n = 1$). For the former, we obtain the following distribution

$$p(y_n|\mathbf{X}, \mathbf{t}, \mathbf{z}, \mathbf{m}) \propto \begin{cases} \mathcal{N}_{y_n}(m_n, 1)\delta(y_n < -1/2) & \text{for} \quad t_n = -1, z_n = 0. \\ \mathcal{N}_{y_n}(m_n, 1)\delta(y_n > 1/2) & \text{for} \quad t_n = 1, z_n = 0 \end{cases}$$

This is a Gaussian either truncated above at $-1/2$ or below at $1/2$. When $z_n = 1$, we marginalise the unobserved t_n (via $p(z_n = 1|y_n) = \sum_{t_n} p(t_n|y_n)p(z_n = 1|t_n)$) resulting in the following distribution

$$p(y_n|m_n, z_n = 1) \propto \gamma_-\delta(y_n < -1/2)\mathcal{N}_{y_n}(m_n, 1) + \gamma_+\delta(y_n > 1/2)\mathcal{N}_{y_n}(m_n, 1)$$

where $\gamma_- = p(z_n = 1|t_n = -1)$ (likewise for γ_+). Defining the pdf of a Gaussian truncated below at a as $\mathcal{N}_y^{>a}(m, 1) = [\Phi(m - a)]^{-1}\delta(y > a)\mathcal{N}_y(m, 1)$ and one truncated above at a as $\mathcal{N}_y^{<a}(m, 1) = [\Phi(a - m)]^{-1}\delta(y < a)\mathcal{N}_y(m, 1)$ we can re-write the conditional distribution as

$$Z^{-1}[\gamma_-\Phi(-1/2 - m_n)\mathcal{N}_{y_n}^{<1/2}(m_n, 1) + \gamma_+\Phi(m_n - 1/2)\mathcal{N}_{y_n}^{>1/2}(m_n, 1)]$$

where $Z = \gamma_-\Phi(-1/2-m_n)+\gamma_+\Phi(m_n-1/2)$. This is a mixture of two truncated Gaussians, and is visualised in Figure 1. These three distributions are all that is required for the Gibbs sampler and sampling from each of them is straightfoward. However, in practice an approximation is likely to be more computationally appealing and so the authors embed the NCNM into a sparse scheme based on the informative vector machine (IVM).

3. Probit GP

Recently, Girolami and Rogers (2006) showed that exact Bayesian inference is possible in binary and multi-class GP classification through augmenting a GP classifier with Gaussian latent variables and a probit likelihood function. In the binary case, the only difference with the model described above is in the choise of noise model which, in this case is defined

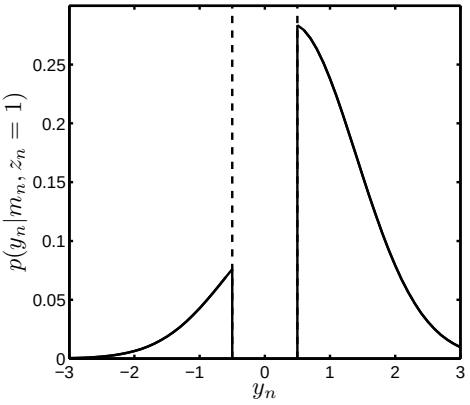


Figure 1: Example conditional posterior for y_n in the binary NCNM. In this case, $m_n = 0.4$ and we see how the posterior pushes y_n outside the null region. The effect m_n has on the mixture weights is obvious. As m_n is close to the right hand edge of the null region, the right hand Gaussian is weighted much higher than the left hand one.

as $p(t_n = 1|y_n) = \delta(y_n > 0), p(t_n = -1|y_n) = \delta(y_n < 0)$. Gibbs sampler updates for y_n are therefore

$$p(y_n|m_n, t_n) \propto \begin{cases} \delta(y_n < 0)\mathcal{N}_{y_n}(m_n, 1) & \text{for} \quad t_n = -1 \\ \delta(y_n > 0)\mathcal{N}_{y_n}(m_n, 1) & \text{for} \quad t_n = 1. \end{cases}$$

The similarity between this and the NCNM updates (when $z_n = 0$) is obvious. The only difference is in the truncation points of the Gaussians and as such, the probit model could be viewed as a special case of the NCNM where $a = 0$ (recall that originally, the width of the null region was defined by a). As an aside, it is intuitive to see what happens if unlabeled data is added to the probit model. Following the same notation as before, we find that

$$p(y_n|z_n = 1, m_n) = Z^{-1}[\gamma_- \Phi(-m)\mathcal{N}_{y_n}^{<0}(m_n, 1) + \gamma_+ \Phi(m)\mathcal{N}_{y_n}^{>0}(m_n, 1)].$$

Where $Z = \gamma_- \Phi(-m) + \gamma_+ \Phi(m)$. Under the standard assumption that $\gamma_- = \gamma_+$, this reduces to the prior $p(y_n|m_n)$ and it is obvious why adding unlabeled data makes no difference in the standard GP framework.

3.1 From binary to multi-class

In the previous section, we described the binary probit GP and its similarity (and crucial difference) to the NCNM. The multi-class classification scheme presented in Girolami and Rogers (2006) has two major advantages compared to other multi-class GP schemes. Firstly, exact inference can be achieved via Gibbs sampling based MCMC and secondly, the scaling

is linear with respect to $\mathcal{K}$, the total number of classes. It is defined as follows

$$\begin{aligned}\mathbf{m}_k &\sim GP(\mathbf{0}, \mathbf{C})\\ \mathbf{y}_k &\sim \mathcal{N}_{\mathbf{y}_k}(\mathbf{m}_k, \mathbf{I}_N)\\ p(t_{nk} = 1|\mathbf{y}_n) &= \delta(y_{nk} > y_{ni} \forall i \neq k)\end{aligned}$$

where $k = 1 \dots K$ — the total number of classes, $\mathbf{M}$ and $\mathbf{Y}$ are both $N \times K$ matrices that can either be decomposed by columns $\mathbf{M} = [\mathbf{m}_1, \dots, \mathbf{m}_k, \dots, \mathbf{m}_K]$ or rows $\mathbf{M} = [\mathbf{m}_1, \dots, \mathbf{m}_n, \dots, \mathbf{m}_N]^T$ where $\mathbf{m}_n$ denotes a $K \times 1$ vector and $\mathbf{m}_k$ a $N \times 1$ vector (resp. for $\mathbf{Y}, \mathbf{y}_n, \mathbf{y}_k$). Additionally, $\mathbf{t}_n$ is now a $1 \times K$ binary vector with a one corresponding to the particular class assigned to this point and zeros elsewhere. Note that in our notation we have assumed that there is one global covariance matrix for all of the K GPs, relaxing this assumption is straightforward. The Gibbs distribution for $\mathbf{y}_n$ is given by

$$p(\mathbf{y}_n|\mathbf{m}_n, \mathbf{t}_n) = \sum_{k=1}^{K} \mathcal{N}_{\mathbf{y}_n}(\mathbf{m}_n, \mathbf{I}_K)\delta(y_{nk} > y_{ni} \forall i \neq k)\delta(t_{nk} = 1),$$

which is a Gaussian located at $\mathbf{m}_n$ truncated such that the component corresponding to the class label is largest. It is possible to visualise this truncation for the case $K = 3$ as can be seen in Figure 2(a) where the sphere represents an iso-contour of a Gaussian with mean zero and the planes represent the boundaries between the truncation regions. An example of a dataset drawn from the prior defined by this model can be seen in Figure 2(c).

As we have seen, the inclusion of a null region can give unlabeled points some influence in the position of our decision boundary. We propose the construction of a null region in the multi-class setting through use of the following modification of the multi-class probit likelihood:

$$p(\mathbf{t}_n|\mathbf{y}_n) = \begin{cases} \delta(y_{nk} > y_{ni} + \epsilon \ \forall i \neq k) & \text{for} \quad t_{nk} = 1 \\ 1 - \sum_k \delta(y_{nk} > y_{ni} + \epsilon \ \forall i \neq k) & \text{for} \quad \sum_k t_{nk} = 0 \end{cases}$$

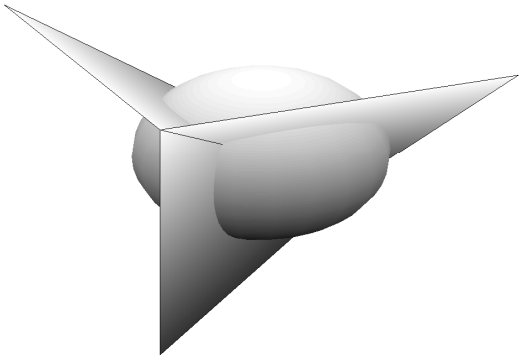
where ϵ is a parameter that controls the width of the null region and points in the null region are characterised by having no assigned label ($\sum_k t_{nk} = 0$). A visualisation of the truncation defined by this model can be seen in Figure 2(b) where the sphere has now been broken into segments. The null region is clearly visible. We denote this scheme as ϵ-truncation. An example dataset drawn from this prior can be seen in Figure 2(d).

Introducing a new set of variables, z_n as before, we can formulate the conditional distributions for $\mathbf{y}_n$

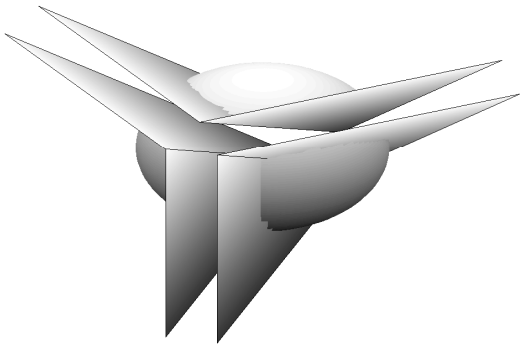
$$p(\mathbf{y}_n|z_n, \mathbf{m}_n, \mathbf{t}_n) \propto \begin{cases} \delta(y_{nk} > y_{ni} + \epsilon \ \forall i \neq k)\mathcal{N}_{\mathbf{y}_n}(\mathbf{m}_n, \mathbf{I}_K) & \text{for} \quad z_n = 0, t_{nk} = 1 \\ \sum_k \gamma_k \delta(y_{nk} > y_{ni} + \epsilon \ \forall i \neq k)\mathcal{N}_{\mathbf{y}_n}(\mathbf{m}_n, \mathbf{I}_K) & \text{for} \quad z_n = 1 \end{cases}$$

where $\gamma_k = p(z_n = 1|t_{nk} = 1)$. As in the binary case, this is a truncated Gaussian when the label is observed and a mixture of truncated Gaussians when it is not. Denoting by $\mathcal{N}^{k,\epsilon}_{\mathbf{y}_n}(\mathbf{m}_n, \mathbf{I}_K) = Z^{-1}_{nk}\delta(y_{nk} > y_{ni} + \epsilon \forall i \neq k)\mathcal{N}_{\mathbf{y}_n}(\mathbf{m}_n, \mathbf{I}_K)$ a Gaussian truncated such that the kth component is the largest, where $Z_{nk} = E_{p(u)}[\prod_{j \neq k} \Phi(u + m_{nk} - m_{nj} - \epsilon)]$, $u \sim \mathcal{N}(0, 1)$ (for details, see appendix A), we can normalise this conditional distribution to give

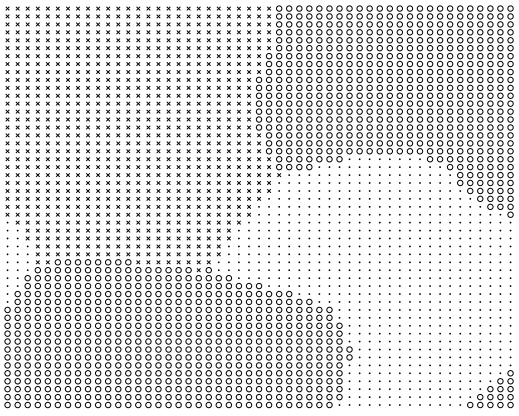
$$p(\mathbf{y}_n|z_n = 1, \mathbf{m}_n) = \frac{\sum_{k=1}^{K} \gamma_k Z_{nk} \mathcal{N}^{k,\epsilon}_{\mathbf{y}_n}(\mathbf{m}_n, 1)}{\sum_{k'} \gamma_{k'} Z_{nk'}}.$$



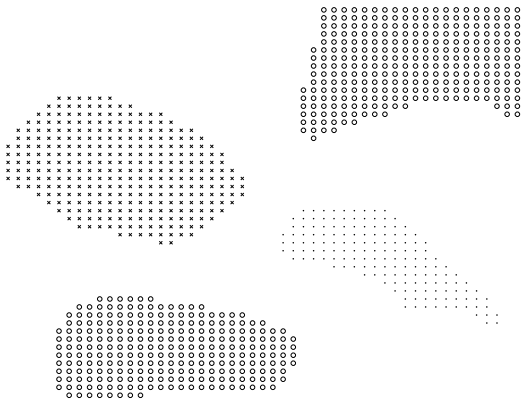
(a) A visualisation of the truncation caused by the standard multi-class probit model



(b) A visualisation of the truncation caused by the multi-class probit model with a null region



(c) Example 3 class dataset drawn from the prior of the standard probit GP. Data points are uniformly spaced and the function has been drawn from a GP prior with an rbf covariance function.



(d) Example 3 class dataset drawn from the prior of the probit GP with null category.

Figure 2: Visualisation of the addition of a null region to the probit GP.

This is all we need to implement a Gibbs sampler. To summarise, the required distributions are

$$\begin{aligned}
\mathbf{m}_k|\mathbf{y}_k &\sim \mathcal{N}_{\mathbf{m}}(\Sigma\mathbf{y}_k, \Sigma) \\
\mathbf{y}_n|z_n = 0, \mathbf{m}_n, \mathbf{t}_n &\sim \sum_{k=1}^{K} \mathcal{N}^{k,\epsilon}_{\mathbf{y}_n}(\mathbf{m}_n, \mathbf{I})\delta(t_{nk} = 1) \\
\mathbf{y}_n|z_n = 1, \mathbf{m}_n &\sim \frac{\sum_{k=1}^{K} \gamma_k Z_{nk} \mathcal{N}^{k,\epsilon}_{\mathbf{y}_n}(\mathbf{m}_n, 1)}{\sum_{k'} \gamma_{k'} Z_{nk'}}
\end{aligned}$$

where $\Sigma = \mathbf{C}(\mathbf{I} + \mathbf{C})^{-1}$ as before.

3.2 Making Predictions

Following Girolami and Rogers (2006), we can obtain the predictive distribution for a new point by marginalising the GP variables, $\mathbf{M}$

$$P(t_{new} = k|\mathbf{x}_{new}, \mathbf{X}, \mathbf{t}, \epsilon) = \int P(t_{new}|\mathbf{m}^{new}, \epsilon)p(\mathbf{m}^{new}|\mathbf{x}_{new}, \mathbf{X}, \mathbf{t})d\mathbf{m}^{new}.$$

Samples from the Gibbs sampler can be used to obtain a Monte Carlo estimate of the above expectation. For each sampled $\mathbf{Y}$, an additional set of samples $m_k^{new,s}$ is drawn from $\mathcal{N}_m(\mu_k^{new}, \sigma_{k,new}^2)$ where $\mu_k^{new} = \mathbf{y}_k^T(\mathbf{I}+C)^{-1}\mathbf{C}^{new}$ and $\sigma_{k,new}^2 = c^{new} - \mathbf{C}^{new}(\mathbf{I}+\mathbf{C})^{-1}\mathbf{C}^{new}$, where $\mathbf{C}^{new}$ is the covariance function evaluated between the new point and the training points and c^{new} is the covariance function evaluated at just $\mathbf{x}^{new}$. The predictive distribution is then given by

$$\frac{1}{N_{samps}} \sum_{s=1}^{N_{samps}} E_{p(u)} \left\{ \prod_{j \neq k} \Phi(u + m_k^{new,s} - m_j^{new,s} - \epsilon) \right\}.$$

However, as mentioned in Lawrence and Jordan (2006), we have a finite probability of our new point belonging to the null class ($1 - \sum_k p(t_{new} = k|\mathbf{x}_{new}, \mathbf{X}, \mathbf{t}, \epsilon) > 0$). We can overcome this problem, as in Lawrence and Jordan (2006), by the reasonable assumption that $z_{new} = 1$ — i.e. that the label for the new point has not been observed. In this case, by definition, $p(z_{new} = 1|t_{new} = 0) = 0$ and so there is zero probability of the new data point belonging to the null class. We can therefore re-normalise our predictive probabilities so that they sum to one over the K legitimate classes.

A sampling scheme such as this, whilst providing samples from the true posterior, is not very appealing from a computational point of view, motivating the creation of a suitable approximation.

4. A Variational Approximation

As in Girolami and Rogers (2006), it is straightforward to obtain a variational Bayes approximation (Beal (2003); Jordan et al. (1999)) in which the posterior over $\mathbf{Y}$ and $\mathbf{M}$ is approximated through a factored posterior $p(\mathbf{M}, \mathbf{Y}|\mathbf{X}, \mathbf{t}, \mathbf{z}, \epsilon) \approx \prod_i Q(\mathbf{\Theta}_i) = Q(\mathbf{Y})Q(\mathbf{M})$. Applying Jensen's inequality to the marginal likelihood, results in the following familiar lower bound $p(\mathbf{t}|\mathbf{X}, \mathbf{Y}, \mathbf{M}, \mathbf{z}, \epsilon) \geq E_{Q(\mathbf{\Theta})}\{\log p(\mathbf{t}, \mathbf{\Theta}|\mathbf{X}, \mathbf{t}, \mathbf{z}, \epsilon)\} - E_{Q(\mathbf{\Theta})}\{\log Q(\mathbf{\Theta})\}$ that is minimised by distributions of the form $Q(\mathbf{\Theta}_i) \propto \exp(E_{Q(\mathbf{\Theta}/\mathbf{\Theta}_i)}\{\log p(\mathbf{t}, \mathbf{M}, \mathbf{Y}|\mathbf{X}, \mathbf{z}, \epsilon)\})$. The form of $Q(\mathbf{M})$ is as given in Girolami and Rogers (2006) and is

$$Q(\mathbf{M}) = \prod_{k=1}^{K} Q(\mathbf{m}_k) = \prod_{k=1}^{K} \mathcal{N}_{\mathbf{m}_k}(\widetilde{\mathbf{m}}_k, \Sigma_k)$$

where $\widetilde{a} = E_{Q(a)}(a)$ and $\widetilde{\mathbf{m}}_k = \Sigma\widetilde{\mathbf{y}}_k$ and $\Sigma = \mathbf{C}(\mathbf{I} + \mathbf{C})^{-1}$. $Q(\mathbf{Y})$ is more interesting and can be factored into the terms corresponding to labeled training examples $n_l = 1 \ldots L$ and

unlabeled $n_u = 1 \dots U$. For the labeled data points, $Q(\mathbf{y}_{n_l}) = \sum_k \mathcal{N}^{k,\epsilon}_{\mathbf{y}_{n_l}}(\mathbf{m}_{n_l}, \mathbf{I})\delta(t_{nk} = 1)$, giving, if $t_{n_l i} = 1$ (i.e. $\mathbf{x}_{n_l}$ belongs to class i),

$$\begin{aligned}
\widetilde{y}_{n_l k} &= \widetilde{m}_{n_l k} - Z^{-1}_{n_l} E_{p(u)} \left\{ \mathcal{N}_u(\widetilde{m}_{n_l k} - \widetilde{m}_{n_l i} + \epsilon, 1) \prod_{j \neq i,k} \Phi(u + \widetilde{m}_{n_l i} - \widetilde{m}_{n_l j} - \epsilon) \right\} \\
&\quad \text{for } k \neq i \text{ and} \\
\widetilde{y}_{n_l i} &= \widetilde{m}_{n_l i} + \sum_{k \neq i} (\widetilde{m}_{n_l k} - \widetilde{y}_{n_l k})
\end{aligned}$$

(for details, see appendix B). For the unlabeled data, we again marginalise over the unobserved $\mathbf{t}_{n_u}$ and obtain

$$Q(\mathbf{y}_{n_u}) = \frac{\sum_k \gamma_k Z_{n_u k} \mathcal{N}^{k,\epsilon}_{\mathbf{y}_{n_u}}(\widetilde{\mathbf{m}}_{n_u}, \mathbf{I})}{\sum_{k'} \gamma_{k'} Z_{n_u k'}}$$

which is a mixture of truncated Gaussians as before. $\widetilde{\mathbf{y}}_{n_u}$ is then given by a weighted combination of the expectations of $\mathbf{y}_{n_u}$ under each of the k truncations which can be calculated as described above. It is interesting to compare this variational approximation with the Gibbs sampler. It is clear that whilst the Gibbs sampler will always sample a value of $\mathbf{y}$ that will move the point out of the null region (or, more accurately, move the null region away from the point), there is no guarantee that $\widetilde{\mathbf{y}}$ will not be inside the null region. For $\mathbf{m}$ right in the centre of the null region, the Gibbs sampler will jump to one side or the other whilst the variational approximation will not move. We will discuss this point further when we compare the two algorithms in the experimental section.

4.1 Variational Predictions

To obtain the predictive distributions, we first marginalise the GP variables to give

$$p(\mathbf{y}^{new}|\mathbf{x}_{new}, \mathbf{X}, \mathbf{t}, \epsilon) = \prod_{k=1}^{K} \mathcal{N}_{y_k^{new}}(\widetilde{m}_k^{new}, \widetilde{v}_k^{new})$$

where $\widetilde{v}_k^{new} = \sqrt{1 + \widetilde{\sigma^2}_k^{new}}$. Applying the ϵ-truncation to this distribution, yields

$$p(t_{new} = k|\mathbf{x}_{new}, \mathbf{X}, \mathbf{t}) = E_{p(u)} \left\{ \prod_{j \neq k} \left(\frac{1}{\widetilde{v}_j^{new}} [u\widetilde{v}_k^{new} + \widetilde{m}_k^{new} - \widetilde{m}_j^{new} - \epsilon] \right) \right\}.$$

As previously, we will need to re-normalise the predictive distributions over the K legitimate classes due to the constraint $p(z_{new} = 1|t_{new} = 0) = 0$.

5. Experimental Results

5.1 Illustrative Example

We will start with a toy example for ease of visualisation. Data are sampled from three Gaussians with means $[-3, 0]^T, [0, 0]^T, [3, 0]^T$ and a shared covariance matrix with an identity on

the diagonal and 0.9 on the off-diagonal elements. Labels are removed from points whose squared euclidean distance from the mean of the distribution from which they were drawn is greater than 0.75. For both the variational approximation and the Gibbs sampling scheme, a value of $\epsilon = 1$ is used with an RBF covariance function $C(\mathbf{x}_i, \mathbf{x}_j) = \exp(-s||\mathbf{x}_i - \mathbf{x}_j||^2)$ with parameter $s = 0.5$. However, as described in Girolami and Rogers (2006), the kernel parameter (or parameters) could be inferred during the training process. Additionally, ϵ could be treated as an additional model parameter. However, following Lawrence and Jordan (2006), we fix it to a constant value (in all experiments, $\epsilon = 1$) and let the GP handle the overall scale. The decision boundaries provided by a classifier trained only on the labeled data and those for the semi-supervised approach can be seen in Figure 3. It is clear that the semi-supervised approach is able to use the information present in the unlabeled data to build a more accurate classifier.

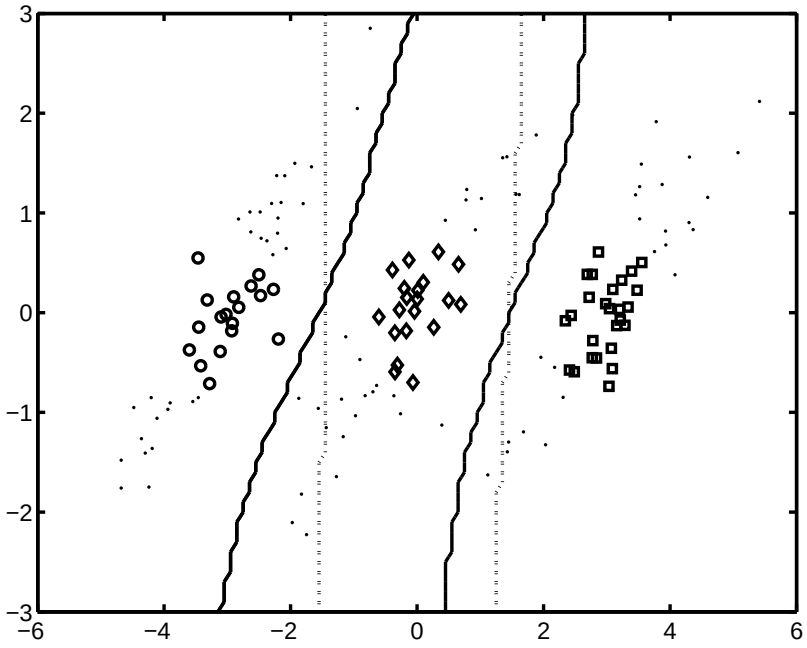
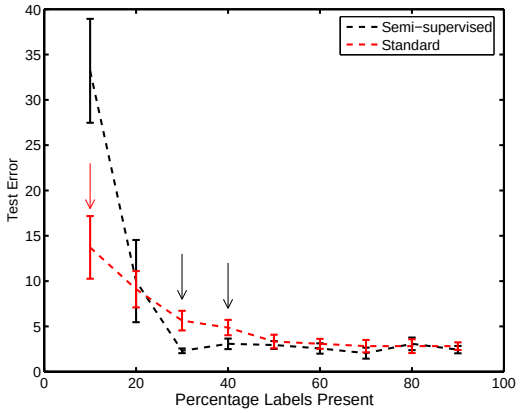


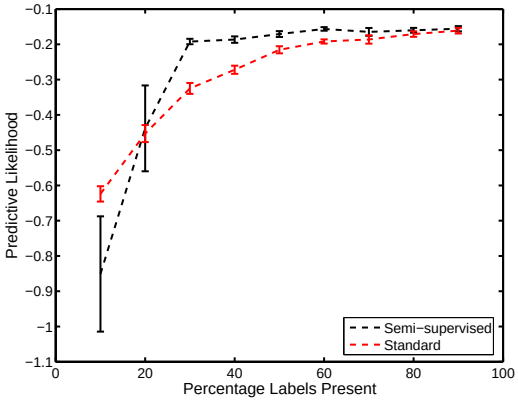
Figure 3: Example semi-supervised problem. Labeled data from three classes is shown as black squares, circles and diamonds. Unlabeled data is shown as grey dots. The decision boundaries without the unlabeled data are shown as dotted lines. When the labeled data is included, the decision boundary is shown as a solid line. The classifier with the unlabeled data more accurately reflects the structure present in the data. There is no visual difference between the Variational approximation and the solution from the Gibbs sampler.

5.2 Wine Dataset

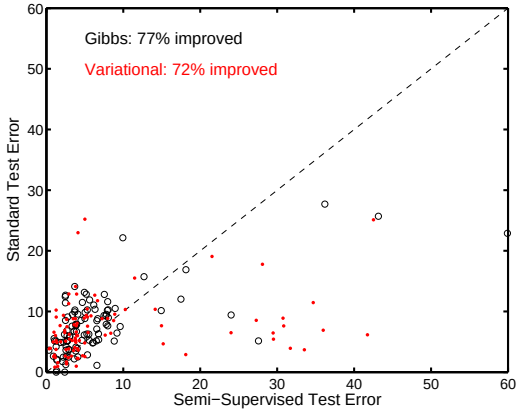
We now turn our attention to the wine dataset from the UCI machine learning repository D.J. Newman and Merz (1998). The dataset consists of 178 instances spread reasonably



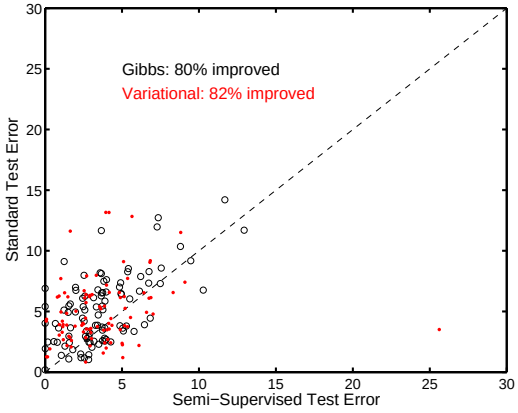
(a) Effect of varying percentage of training points labeled on classification performance for variational approximation. Error bars show standard error over 10 re-samplings, arrows show points where difference is significant at 5% under a paired t-test.



(b) Effect of varying percentage of training points labeled on predictive likelihood for Gibbs sampler. Error bars show standard error over 10 re-samplings and the semi-supervised approach is significantly better at all percentages between 30 and 80 inclusive.



(c) Scatter plot showing comparison of performance with and without unlabeled data at 20% labeled. Circles represent results using Gibbs samples, dots the variational approximation.



(d) As (b), but at 30% labeled.

Figure 4: Effect of varying the proportion of training points left unlabeled for the wine dataset.

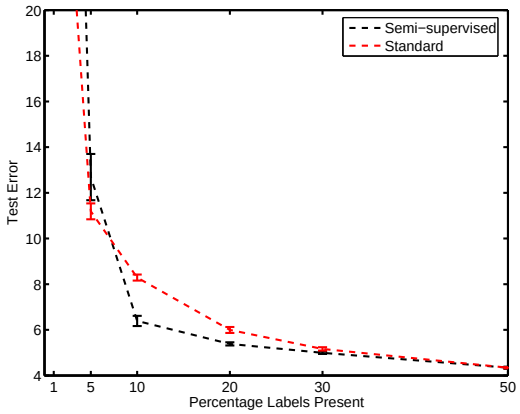
evenly across three classes. Twelve features are measured and, due to the large difference in their scales, we re-scale them by subtracting the mean and dividing by the standard deviation of the training data. The dataset is randomly partitioned into 100 training and 78 testing points and we investigate the effect of removing different proportions of the labels of the training data. We use an RBF kernel and fix both the kernel parameter $s = 0.1$ and the width of the null region, $\epsilon = 1$. Figure 4(a) shows how the test error varies as the proportion of labeled training data is increased. We see that at the lowest percentages of

labeled data, the performance for the semi-supervised approach is significantly inferior to the standard classifier and performances for both approaches are characterised by a very large variance. As the percentage is increased, the performance of the semi-supervised approach improves rapidly and at 30% and 40% is significantly better than the performance with the labeled data alone. However, closer analysis of these results suggests that simply comparing the means under the different approaches does not provide the full story. In Figures 4(c) and 4(d) we compare the error rates achieved for both the variational approximation and the Gibbs sampler over 100 re-samplings of the data (to ease visualisation, a small quantity of random *jitter* has been added to the points — without it, many points lie on top of one another due to the small dataset size). The two plots correspond to 20% labeled (where the mean performances are indistinguishable) and 30% labeled where the semi-supervised approach is significantly better. Also shown on the plots are the percentage of cases in which semi-supervised learning gave an improvement. We see that in both cases, the performance under the majority of partitions was improved when unlabeled data was added. However, particularly in the 20% labeled case, the improvements are reasonably modest when compared to the few cases where the performance is substantially reduced. Although this is only one dataset and empirical performance is likely to vary substantially across different domains, this is an important point and will be discussed further in the conclusions. In addition to the test error, as we have a probabilistic classifier, we can also monitor the predictive likelihood. Figure 4(b) shows how the average predictive likelihood varies as the percentage of data labeled is increased. The semi-supervised approach has significantly higher predictive likelihood than the standard approach for all percentages between 30 and 80 inclusive. This suggests that whilst there are only modest increases in performance with respect to test error, there is a larger difference in the predictive likelihood — the unlabeled data is providing increased certainty in predictions. In both the test error and predictive likelihood, there is no significant difference in the performance of the Gibbs sampler and variational approximation. This generally agrees with the results for standard classification presented in Girolami and Rogers (2006).

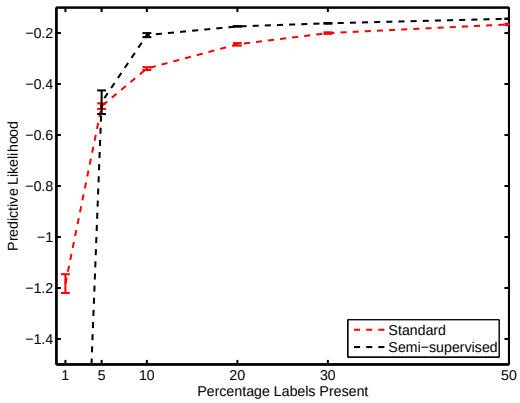
5.3 USPS Digits

Finally, we turn our attention to a much larger problem. The USPS digits dataset (Hull, 1994)[1]consists of 9298 images of the digits 0-9 split into equal training and test sets. Each image is described by 256 features (corresponding to the 16×16 grey level pixel intensities). A discriminative classifier is particularly appropriate in examples such as this as building a model of $p(\mathbf{x})$ in such a high dimensional space would be incredibly difficult. In all experiments, an RBF kernel was used with $s = 0.01$. As in the previous section, labels were removed with probability ranging between 0.01 and 0.5 and at each probability, the experiment was repeated 10 times. Figures 5(a) and 5(b) show the test error and predictive likelihood for the semi-supervised and standard approached as the percentage of labeled points is increased. As in the previous example, we see that at very low percentages, the performance of the semi-supervised technique is much worse than using the labeled data alone. However, as the proportion of labels is increased, the performance of the semi-

1. Particular version used here is available from `http://www.gaussianprocess.org/gpml/data/`



(a) Effect of varying the percentage of labeled points on test error in the 10-class USPS handwritten digits dataset



(b) Effect of varying the percentage of labeled points on predictive likelihood in the 10-class USPS handwritten digits dataset

Figure 5: Performance on large-scale USPS dataset

supervised approach improves and for 10 and 20 percent, there is a considerable advantage in including the unlabeled data.

6. Conclusions

In this paper, we have shown how the null category noise model (NCNM), as introduced into the GP literature by Lawrence and Jordan (2006) can be incorporated into a multi-class setting through the multinomial probit GP of Girolami and Rogers (2006) and thus allow semi-supervised learning to take place in a multi-class, discriminative setting. Semi-supervised learning in a discriminative setting is desirable as there is no need to make often computationally costly approximations of the data density $p(\mathbf{x})$. One of the major advantages of the probit GP is that it allows exact inference through MCMC. To this end, we present a Gibbs sampler and also provide a more computationally attractive variational approximation.

Answers to such questions as to whether or not semi-supervised learning will offer great improvements over standard approaches and whether generative or discriminative approaches will prove more useful, while important, are certainly not the aim of this paper. This explains the relatively short experimental section. However, some interesting questions arise from our results. Particularly, at (relatively) small percentages of labeled samples (a scenario which is common and motivates the use of semi-supervised approaches) the very similar mean performances mask considerable differences between the performances of the standard and semi-supervised approaches. For example, in our experiments on the wine dataset, at 20% labeled data the majority of partitions are improved by the addition of unlabeled data. However, there is no significant difference between the mean performances of the semi-supervised and standard approaches due to the small number of partitions where performance is substantially hampered by the unlabeled points — incusion of unlabeled data is risky. It would seem reasonable to assume that in order to gain benefit from in-

cluding unlabeled data, the distribution of the labeled data needs to be biased with respect to the true distribution from which the data was generated. The addition of more data (be it labeled or unlabeled) has the potential to help correct this bias. However, large decreases in performance with the inclusion of unlabeled data could be due to the distribution of the labeled data being *too* different from the true distribution, effectively causing unlabeled points to sit on the wrong side of the decision boundary. It would be interesting to see whether it would be possible to gain an insight, given a particular dataset, into likely improvements/reductions in classification performance without the aid of testing and validation data. However, it is difficult to see how this would be possible without creating a model of $p(\mathbf{x})$ - exactly what we are trying to avoid in the discriminative framework.

Whilst any improvements found in the test error were modest, we see a far more significant improvement in the predictive likelihood. This suggests that, over a certain percentage of labeled points, the addition of unlabeled data helps to make the predictions more certain even if no significant difference in test error is observed. Additionally, it is possible that the rather modest improvements found were in part due to the method of randomly removing labels. Such a process doesn't reflect the likely real scenario of the distribution of labeled data systematically deviating from the underlying data distribution. Whilst this makes the classification task more difficult it may mean that semi-supervised results look more promising as the performance of algorithms trained only on the labeled data is likely to diminish.

One of the major drawbacks of using a GP for classification in many practical settings is the dominant $\mathcal{O}(N^3)$ scaling resulting in the inversion of the covariance matrix. Producing a sparse solution to the classification problem (i.e. using a subset of training points $S << N$) is one way to overcome this problem, for example the IVM (Lawrence et al., 2005) or the sparse online GPs of Csato and Opper (2002). In Girolami and Rogers (2006), we showed how the multi-class probit GP could be made sparse through IVM-like updates. It is interesting to consider how such an algorithm would behave in the presence of unlabeled data. For example, what proportion of unlabeled data do we have in the final solution and when choosing a new point, should we make allowances for the potential additional information that is available from a labeled data point. These are interesting questions for future research.

Finally, we have not made any mention of suitable values for ϵ. The optimum value is likely to be dependent on both the particular dataset being used and the form of the covariance function. For example, given a particular dataset, as ϵ is increased, the GP will require more flexibility to ensure that the closest points to the decision boundary do not invalidate the necessary constraints. This implies that for a given covariance function, different values of ϵ may favour completely different decision functions. An investigation into the role the value of ϵ plays and whether or not it can be learnt from the data is an interesting avenue for future work.

7. Acknowledgments

SR and MG are supported by EPSRC grant EP/CO10620/1 — Stochastic modelling and statistical inference of gene regulatory pathways: integrating multiple sources of data.

References

M. Beal. *Variational Algorithms for approximate Bayesian inference*. PhD thesis, University College London, 2003.

O. Chapelle, J. Weston, and B. Schölkopf. Cluster kernels for semi-supervised learning. In *Advances in Neural Information Processing Systems 15*, 2003.

L. Csato and M. Opper. Sparse online Gaussian processes. *Neural Computation*, 14(2): 641–668, 2002.

C.L. Blake D.J. Newman, S. Hettich and C.J. Merz. UCI repository of machine learning databases, 1998. URL `http://www.ics.uci.edu/~mlearn/MLRepository.html`.

M. Girolami and S. Rogers. Variational bayesian multinomial probit regression with Gaussian process priors. *Neural Computation*, 18:1790–1817, 2006.

J.J. Hull. A database for handwritten text recognition research. *IEEE PAMI*, 16(5):550–554, 1994.

M.I. Jordan, Z. Ghahramani, T.S. Jaakkola, and L.K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 1999.

J.A. Lasserre, C.M. Bishop, and T.P. Minka. Principled hybrids of generative and discriminative models. In *Proceedings 2006 IEEE Conference on Computer Vision and Pattern Recognition, New York*, 2006.

N.D. Lawrence and M.I. Jordan. Gaussian processes and the null-category noise model. In O. Chapelle, B. Schölkopf, and A. Zien, editors, *Semi-supervised Learning*, chapter 8. MIT Press, 2006.

N.D. Lawrence, J.C. Platt, and M.I. Jordan. Extensions of the informative vector machine. In J. Winkler, N.D. Lawrence, and M. Niranjan, editors, *Deterministic and Statistical Methods in Machine Learning*. Springer-Verlag, 2005.

M. Seeger. Covariance kernels from bayesian generative models. In *Advances in Neural Information Processing Systems 14*, 2001.

C.K. Williams and D. Barber. Bayesian classification with Gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1998.

X. Zhu, J. Lafferty, and Z. Ghahramani. Semi-supervised learning - from Gaussian fields to Gaussian processes. Technical report, Carnegie Mellon University, 2003.

Appendix A. Normlisation constant of conically ϵ-truncated Gaussian

The normalisation constant can be calculated as follows

$$\begin{aligned} Z_{ni} &= \int_{y_{ni}=\infty}^{\infty} \mathcal{N}_{y_{ni}}(m_{ni},1) \prod_{k\neq i} \int_{y_{nk}=-\infty}^{\infty} \mathcal{N}_{y_{nk}}(m_{nk},1)\delta(y_{ni} > y_{nk}+\epsilon) dy_{nk} dy_{ni} \\ &= \int_{y_{ni}=\infty}^{\infty} \mathcal{N}_{y_{ni}}(m_{ni},1) \prod_{k\neq i} \int_{y_{nk}-\infty}^{y_{ni}-\epsilon} \mathcal{N}_{y_{nk}}(m_{nk},1) dy_{nk} dy_{ni} \\ &= \int_{y_{ni}=\infty}^{\infty} \mathcal{N}_{y_{ni}}(m_{ni},1) \prod_{k\neq i} \Phi(y_{ni}-m_{nk}-\epsilon) dy_{ni}. \end{aligned}$$

Making the substitution $u = y_{ni} - m_{ni}$ leaves us with

$$\begin{aligned} Z_{ni} &= \int_{u=\infty}^{\infty} \mathcal{N}_u(0,1) \prod_{k\neq i} \Phi(u + m_{ni} - m_{nk} - \epsilon) du \\ &= E_{p(u)}\left[\prod_{k\neq i} \Phi(u + m_{ni} - m_{nk} - \epsilon) \right] \end{aligned}$$

where $p(u) = \mathcal{N}_u(0,1)$.

Appendix B. Expected value of a conically ϵ-truncated Gaussian

Firstly, assume that $t_{ni} = 1$, i.e. the nth point belongs to class i. So, starting with $\tilde{y}_{nk}\ \forall k \neq i$, we have

$$\tilde{y}_{nk} = Z_n^{-1} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y_{nk} \mathcal{N}_{y_{ni}}(\tilde{m}_{ni},1) \prod_{j\neq i} \mathcal{N}_{y_{nj}}(\tilde{m}_{nj},1)\delta(y_{nj} < y_{ni} - \epsilon) dy_{nj} dy_{ni}.$$

$Z_n = E_{p(u)}\left\{\prod_{j\neq i} \Phi(u + m_{ni} - m_{nj} - \epsilon)\right\}$ as we have shown previously. Expanding this expression gives

$$\tilde{y}_{nk} = Z_n^{-1} \int_{-\infty}^{\infty} \int_{y_{nk}=-\infty}^{y_{ni}-\epsilon} y_{nk} \mathcal{N}_{y_{nk}}(\tilde{m}_{nk},1) \prod_{j\neq i,k} \mathcal{N}_{y_{ni}}(\tilde{m}_{ni},1) \Phi(y_{ni} - \tilde{m}_{nj} - \epsilon) dy_{ni} dy_{nk}.$$

Applying the substitution $u = y_{nk} - \tilde{m}_{nk}$ results in

$$\begin{aligned} \tilde{y}_{nk} &= Z_n^{-1} \int_{-\infty}^{\infty} \int_{u=-\infty}^{y_{ni}-\epsilon-\tilde{m}_{nk}} (u + \tilde{m}_{nk}) \mathcal{N}_u(0,1) \prod_{j\neq i,k} \mathcal{N}_{y_{ni}}(\tilde{m}_{ni},1) \Phi(y_{ni} - \tilde{m}_{nj} - \epsilon) dy_{ni} du \\ &= \tilde{m}_{nk} - Z_n^{-1} E_{p(u)}\left\{ \mathcal{N}_u(\tilde{m}_{nk} - \tilde{m}_{ni} + \epsilon, 1) \prod_{j\neq i,k} \Phi(u + \tilde{m}_{ni} - \tilde{m}_{nj} - \epsilon) \right\} \end{aligned}$$

where we have used the results for a standard one-dimensional truncated Gaussian. For $\tilde{y}_{ni}$, we have

$$\tilde{y}_{ni} = Z_n^{-1} \int_{-\infty}^{\infty} y_{ni} \mathcal{N}_{y_{ni}}(\tilde{m}_{ni}, 1) \prod_{j \neq i} \Phi(y_{ni} - \tilde{m}_{nj} - \epsilon) dy_{ni}.$$

Again, substituting $u = y_{ni} - \tilde{m}_{ni}$

$$\begin{aligned} \tilde{y}_{ni} &= Z_n^{-1} \int_{-\infty}^{\infty} (u + \tilde{m}_{ni}) \mathcal{N}_u(0, 1) \prod_{j \neq i} \Phi(u + \tilde{m}_{ni} - \tilde{m}_{nj} - \epsilon) du \\ &= \tilde{m}_{ni} + Z_n^{-1} E_{p(u)} \left\{ u \prod_{j \neq i} \Phi(u + \tilde{m}_{ni} - \tilde{m}_{nj} - \epsilon) \right\}. \end{aligned}$$

Using the result that for $u \sim \mathcal{N}(0, 1)$ and any differentiable function $g(u)$, $E_{p(u)}[ug(u)] = E_{p(u)}[g'(u)]$ (where $g'(u) = dg(u)/d(u)$), it follows that the expectation for y_{ni} can be calculated from the the expectations for $y_{nk}, k \neq i$ via

$$\tilde{y}_{ni} = \tilde{m}_{ni} + \sum_{k \neq i} \tilde{m}_{nk} - \tilde{y}_{nk}.$$