The use of unlabelled data in predictive modelling

Feng Liang, Sayan Mukherjee and Mike West*
Duke University

June 21, 2006
Original version: May 2005

Abstract

The incorporation of unlabelled data in regression and classification analysis is an increasing focus of the applied statistics/machine learning literatures, with a number of recent examples demonstrating the potential for unlabelled data to contribute to improved predictive accuracy. The statistical basis for this semi-supervised analysis does not appear to have been well-delineated; as a result, the underlying theory and rationale may be under-appreciated, especially by non-statisticians. There is also room for statisticians to become more fully engaged in the vigorous research in this important area of intersection of the statistical and computer sciences. Much of the theoretical work in the literature has focused, for example, on geometric and structural properties of the unlabeled data in the context of particular algorithms, rather than probabilistic and statistical questions. This paper overviews the fundamental statistical foundations for predictive modelling and the general questions associated with unlabelled data, highlighting the relevance of venerable concepts of sampling design and prior specification. This theory, illustrated with a series of central illustrative examples and two substantial real data analysis, shows precisely when, why and how unlabelled data matter.

Keywords: Bayesian analysis, Bayesian kernel regression, Latent factor models, Mixture models, Predictive distribution, Semi-supervised learning, Unlabelled data.

---

*Feng Liang is Assistant Professor of Statistics & Decision Sciences, Duke University; Sayan Mukherjee is Assistant Professor of Statistics & Decision Sciences, and of Computer Science, and in the Institute for Genome Sciences & Policy, Duke University; Mike West is The Arts & Sciences Professor of Statistics & Decision Sciences, and Professor of Biostatistics & Bioinformatics, Duke University. Contact author: Feng Liang, Box 90251, Duke University, Durham NC 27708 - feng@stat.duke.edu
1 Introduction

Recent interest in the use of so-called unlabelled data in problems of prediction in the machine learning community has generated a growing awareness of the potential for incorporation of ancillary design data in classification and regression problems (Joachims, 1999; Blum and Mitchell, 1998; Szummer and Jaakkola, 2001; Zhu et al., 2003; Belkin et al., 2004; Bennett and Demiriz, 1998). This use of unlabelled data is often referred to as semi-supervised learning. Mainstream probabilistic thinking is relatively under-represented in this active and exciting literature, and the theoretical underpinnings of algorithms that exploit unlabelled data have received scant attention from statistical scientists. Much of the activity is algorithmic and applied. Machine learning examples are typically presented case-by-case, with the semi-supervised analysis usually based on modifications of (fully supervised) optimization algorithms for classification or regression prediction, and with the introduction of additional components of objective functions that tie-in unlabelled samples. Arguments for these additional components are made using a combination of structural and intuitive arguments, including, most recently, asymptotic arguments on the convergence of operators on manifolds (Belkin and Niyogi, 2005; Coifman et al., 2005a,b). There has been some work addressing the theoretical aspects of unlabelled data (Seeger, 2000; Castelli and Cover, 1995; Cozman and Cohen, 2002; Ando and Zhang, 2005) in specific contexts. However, in general, the foundation and rationale for understanding the relevance, and likely effectiveness, of unlabelled data is still not well understood.

For both currently active application areas and to underlie growth and development of the unlabelled data methodology long-term, it is critical that the underlying theoretical basis for the use of unlabelled data be delineated and more broadly understood among statistical and computational scientists. Our goal here is to promote broader awareness and interest among statisticians of the nature and importance of this area. We do this by outlining the conceptual and theoretical bases for the “when, why and how” in regard to the use of unlabelled data, and through a complementary series of illustrations in central statistical modelling contexts as well as empirical examples in two substantive data analysis.

Beginning in Section 2 with an articulation of the basic model framework and discussion of fundamental issues of sampling and design, we discuss the underlying conceptual and theoretical basis for using unlabelled data. This is developed in the Bayesian framework for prediction, in which implications for the incorporation, or otherwise, of unlabelled data in prediction problems becomes transparent. Section 3 provides concrete, illuminating examples in a series of common statistical models. This includes examples in regression, prediction using multivariate normal mixture models, and standard mixture-based classification and discrimination. These are key contexts that connect intimately with some of the major areas of interest in machine
learning, and contexts in which the relevance of unlabelled data is perhaps most transparent and intuitive. These examples serve to highlight the relevance of unlabelled data in standard, central areas of statistics. Section 4 overviews and exemplifies the issues in a class of latent factor regression models, with an empirical illustration in analysis of a benchmark data set of hand-written digit classification. Section 5 concerns our final important context, that of kernel regression; here we link statistical and machine learning approaches, illustrate the theoretical basis for the use of unlabelled data, and provide a further empirical study in discrimination of cancer and normal tissue samples based on gene expression data. We close with summary comments in Section 6.

2 General framework

2.1 Context, goals and models

Interest lies in aspects of the joint distribution of two random quantities, \((y, x)\), and the core prediction problem concerns statements about future values of \(y\) based on observing the corresponding \(x\). Both \(x\) and \(y\) may be multivariate, in general. In standard regression problems, \(y\) is a continuous or discrete univariate response; in problems of classification, \(y\) is discrete, often binary. Using \(p(\cdot)\) as generic notation for probability density functions, all inference problems require understanding aspects of the joint density \(p(y, x|\star)\), where \(\star\) denotes all parameters – to be described below in context – that are needed to fully specify the joint density.

The fundamental problem of prediction – whether it be couched in terms of regression estimation or classification – is framed as follows: at a specified “future” value of \(x\), make statements about the corresponding value of \(y\). Using \(\star\) to denote future values of interest, this implies a directional focus: we want to understand and evaluate, or estimate, \(p(y_*|x_*, D)\) based on all available data and information \(D\).

Statistical models structure the problem in terms of parameters (which may be infinite dimensional in non-parametric models) that represent all uncertain aspects of the joint probability distribution for \((y, x)\). By way of notation, the dominant and generally (our) preferred specification of the joint density is

\[
p(y, x|\phi, \theta) = p(y|x, \phi)p(x|\theta)
\]  

(1)

where the functional forms of the two densities on the right hand side are completely specified by the characterizing parameters \((\phi, \theta)\). The parameters \(\phi\) and \(\theta\) relate explicitly to the conditional for \(y\) given \(x\) and then the marginal for \(x\), respectively. Though \(\phi\) and \(\theta\) are two distinct symbols in notation, they can be structurally dependent in various ways as we will see later. From this joint density, we can also deduce the implied marginal density for \(y\), \(p(y|\phi, \theta)\), and
the implied conditional density $p(x|y, \phi, \theta)$ via the complementary factorization

$$p(y, x|\phi, \theta) = p(x|y, \phi, \theta)p(y|\phi, \theta),$$  

where the full set of parameters $(\phi, \theta)$ may be involved, in complicated ways, in the “retro-

spective” conditional for $x$ given $y$, and the corresponding marginal for $y$.

The conditional density of $y$ given $x$ is essential for prediction, of course, and hence we

center our development on the representation (1), in the knowledge that we can move inter-

changeably between factorizations (1) and (2) as desired.

2.2 Sampling designs

The stochastic model of the data generation process, referred to as the sampling design, leads
to likelihood functions as summaries of the data-based information on $(\phi, \theta)$. Typical sampling
contexts fall into the following categories:

1. Data from the margins:

   - $Y^m = \{y_i^m, i = 1 : k_m\}$ where the $y_i^m \sim p(y|\phi, \theta)$ are independent, and/or
   - $X^m = \{x_i^m, i = (k_m + 1) : (k_m + n_m)\}$ where the $x_i^m \sim p(x|\theta)$ are independent,

   and with $Y^m$ and $X^m$ independent given $(\phi, \theta)$. Having the opportunity to observe
data $Y^m$ provides information on aspects of the full set of parameters $(\phi, \theta)$, while $X^m$
informs on aspects of $\theta$ alone. $X^m$ is the traditional unlabelled data, though the same
term could also be applied to $Y^m$.

2. Full prospective random sampling in which $(Y^p, X^p) = \{(y_i^p, x_i^p); i = 1 : n_p\}$ are
drawn from the full joint distribution $p(y, x|\phi, \theta)$. Here data are paired and provide
information on both $\theta$ and $\phi$. This is a common classification and/or regression design.

3. Data from a prospective design in which the $X^p = \{x_i^p, i = 1 : n_p\}$ values above are
specified in advance by design. Then $X^p$ contains no information about the parameters
and we learn about the parameter $\phi$ (only) through the likelihood based on $(Y^p, X^p)$ that
is the product of components $p(y_i^p|x_i^p, \phi)$ – this is the venerable and perhaps the most
common regression design in applied statistics. In machine learning the term transduc-
tive framework, outlined by Vapnik (1998), has been applied in this setting when the
objective is to make predictions of $y$ on only some specific, pre-specified values of $x$.

4. Data from a typical retrospective design – or case-control design – in which we observe
the outcomes $X^r = \{x_i^r, i = 1 : n_r\}$ at a chosen set of $y$ values $Y^r = \{y_i^r, i = 1 : n_r\}$.
Here too the data are paired, but $Y^r$ provides no information about $(\phi, \theta)$ since the $y$
values are chosen by design. The data in $X^r$ comprise a set of $n_r$ independent random
draws from $p(x|y, \phi, \theta)$ and therefore provide information about $(\phi, \theta)$. 

4
The difference between “prospective” and “retrospective” is whether the observed \( y \) values are random or not. Since most examples we will discuss come from a prospective design, for notational simplicity we will drop the superscript and use \((Y, X)\) to denote \((Y^p, X^p)\). Other sampling schemes arise in statistical design (e.g., matched case-control designs, repeated measurement designs), but the above examples are key and central to much of predictive modelling and to our main goals of explicating the use of unlabelled data. Finally we note that the machine learning community has used the term “sampling” for a somewhat different use, applying it to different factorizations of the joint distribution assuming the data was generated by a full random sampling of \((y, x)\); in that usage, the form (1) is referred to as “diagnostic sampling” and (2) is referred to as “generative sampling” (Cozman and Cohen, 2002).

2.3 Prediction

We observe data \( D \) generated via one or a combination of the sampling designs mentioned above. We aim to predict (estimate, classify) a new case \( y^* \) at a value \( x^* \). The prediction problem is solved from the Bayesian perspective by evaluating the posterior predictive distribution

\[
p(y^*_x, D) = \int \int p(y^*_x, \phi)p(\phi, \theta | x^*_x, D)d\phi d\theta \tag{3}
\]

at the value of the future \( x^* \), where \( p(\phi, \theta | x^*_x, D) \) is the posterior distribution of the parameters given the data and \( x^*_x \). This posterior predictive distribution is the relevant quantity whether \( x^*_x \) is a random draw from \( p(x | \theta) \) or is specified directly. In the former case \( x^*_x \) arises as a sample from \( p(x | \theta) \) and so provides additional information about \( \theta \); then \( p(\phi, \theta | x^*_x, D) \) depends on \( x^*_x \). In the latter case \( x^*_x \) is chosen at a value of interest, often one of a range of values where we aim to explore potential future outcomes, and so provides no additional information; then

\[
p(\phi, \theta | x^*_x, D) = p(\phi, \theta | D). \tag{4}
\]

In any example it is important to be aware of the distinction but, for our development, it is a side issue and we assume the latter case (4) as it simplifies the notation.

Our interest focuses on how \( X^m \) enters in the evaluation of the predictive density in equation (3). All forms of information enter through \( D \), so for \( X^m \) (and any other information) to be relevant in prediction it is necessary that it play a role in defining the posterior \( p(\phi, \theta | D) \). This is the key to understand, if and, how any information in \( D \) impacts the prediction problem.

A relatively general framework has observations on each of \( Y^m, X^m, (Y, X) \) and \((Y^r, X^r)\). Then Bayes’ theorem under a specified prior \( p(\phi, \theta) \) yields

\[
p(\phi, \theta | D) \propto p(\phi, \theta)p(D | \phi, \theta)
\]

with

\[
p(D | \phi, \theta) = p(Y, X | \phi, \theta)p(X^m | \phi, \theta)p(Y^m | \phi, \theta)p(X^r | Y^r, \phi, \theta).
\]
This posterior will depend in complicated ways on all aspects of \( D \), including aspects of the unlabelled data \( X^m \). Investigating this dependence is the key to understanding the relevance and specific potential uses of unlabelled data.

### 2.4 Common framework of regression and classification

For convenience and clarity, we start our discussion in the simple regression/classification context where data arise from a joint random sample \( D = (X, Y) \). Then

\[
p(\phi, \theta | D) \propto p(\phi, \theta)p(Y|X, \phi)p(X|\theta).
\]

For example, we may have a linear or nonlinear regression model for \((y|x, \phi)\) in which \( \phi \) represents the uncertain regression parameters or regression functions.

Now imagine that we have the opportunity to additionally observe or measure some unlabelled data \( X^m \). The modified posterior with \( D = \{Y, X, X^m\} \) is then

\[
p(\phi, \theta | D) \propto p(\phi, \theta)p(Y|X, \phi)p(X|\theta)p(X^m|\theta).
\]

If \( \phi \) and \( \theta \) are independent under the prior, \( p(\phi, \theta) = p(\phi)p(\theta) \), then

\[
p(\phi, \theta | D) = p(\phi|Y, X)p(\theta|X^m, X).
\]

Thus, prior independence leads to posterior independence and the unlabelled data \( X^m \) is irrelevant in learning about \( \phi \), and hence irrelevant in predicting new \( y_* \), if \( \phi \) and \( \theta \) are \textit{a priori} independent. This follows from

\[
p(y_*|x_*, D) = \int \int p(y_*|x_*, \phi)p(\phi, \theta | D)d\phi d\theta = \int p(y_*|x_*, \phi)p(\phi|Y, X)d\phi
\]

by posterior independence.

In other cases, the posterior for \((\theta, \phi)\) may – and often will – involve dependencies. Therefore, additional information generated from marginal data will have an impact on the prediction problem via the integration over the posterior that defines \( p(y_*|x_*, D) \). In the general framework, data from \( Y^m, X^m \), and \((Y^r, X^r)\) will all have an impact on the prediction problem.

It is now evident that, beyond probabilistic/prior dependencies in the Bayesian formulation, any structural relationships between the “regression component” parameters \( \phi \) and the “\( x \)-marginal” component parameters \( \theta \) will also inevitably lead to dependence of predictions on the unlabelled data. How such dependencies arise and what forms they take depend on context.

A standard and conceptually simple example is that of mixture modelling, in which learning about the marginal distribution for \( x \) informs on the relative probabilities of mixture components for the joint distribution for \((y, x)\), and hence influences predictions. This idea-fixing example is developed below as the first of a series of common modelling contexts that illuminate the issues and theoretical framework.
3 Some central modelling contexts and examples

3.1 Non-linear regression prediction using mixtures

A methodologically central example, and one in which the relevance of unlabelled data is transparent, is that of Gaussian mixture modelling for regression. Consider the case of univariate $y$ and multivariate $x$, with joint sampling density

$$p(x, y) = \sum_{i=1}^{m} \pi_i f_i(x, y),$$

where $0 \leq \pi_i \leq 1$, $\sum_i \pi_i = 1$, and the $f_i(x, y)$ are distinct multivariate normal distributions. Transforming the joint distribution (5) to the prospective parametrization in (1) we have

$$p(x | \theta) = \sum_{i=1}^{m} \pi_i f_i(x), \quad p(y | x, \phi) = \sum_{i=1}^{m} w_i(x) f_i(y|x),$$

where $f_i(x)$ and $f_i(y | x)$ are the corresponding marginal and conditional distributions of the multivariate normal $f_i(y, x)$, and

$$w_i(x) = \frac{\pi_i f_i(x)}{\sum_j \pi_j f_j(x)}$$

is the conditional mixing probability evaluated at the conditioning $x$ value.

In terms of the general theory and notation of Section 2 above, the simplest parameter specification has $\phi = \{\pi, \alpha, \beta\}$ and $\theta = \{\pi, \alpha\}$ where $\pi = \{\pi_i : i = 1 : m\}$, $\beta$ is the full set of linear regression coefficients, intercepts and conditional variances in the set of $m$ normal linear models $f_i(y|x)$, and $\alpha$ is the full set of mean vectors and variance matrices of the normal distributions $f_i(x)$. This parametrization makes clear the direct structural dependence of $\phi$ and $\theta$, and hence the deductions from the general theory of Section 2 that unlabelled data will matter in future predictions. This conclusion is evident by inspection. Observing data on the margin $x$ provides direct information on the relative weights $\pi_i$ of the normal components, and hence provides information relevant to predicting future $y$ values. The unlabelled data also of course informs about the other parameters $\alpha$ of the margin for $x$, i.e., the component mean and covariance parameters of the marginal normal mixture $p(x|\theta)$ as well as the component weights. These parameters are also involved in the calculations needed for prediction – through the conditional mixing probabilities $w_i(x*)$ – and so the unlabelled data plays a more intricate role than just advising on the weights.

An illustrative example

A simple illustrative example fixes ideas and shows how unlabelled data may increase predictive accuracy in non-linear regression via a mixture model. Consider two-dimensional data
(y, x) modelled using a three-component Gaussian mixture model in the above framework, setting \( f_i(y, x) \) to be the bivariate normal \( N(\mu_i, \Sigma_i) \), for each \( i = 1 : 3 \). From the following prior on model parameters, we simulated one set of parameters and, given those parameters, drew a sample of 175 observations from the resulting three component Gaussian mixture. Figure 1(a) is a scatter plot of the 175 observations.

The prior,

\[
\begin{align*}
(\pi_1, \pi_2, \pi_3) & \sim \text{Dir}(1/3, 1/3, 1/3), \\
(\mu_i | \Sigma_i) & \sim N(0, \tau \Sigma_i), \ (i = 1, 2, 3), \ \text{with } \tau = 0.2, \\
\Sigma_i & \sim \text{IW}(d, S_0), \ (i = 1, 2, 3), \ \text{with } d = 3 \ \text{and } S_0 = (4/3) I_{2\times 2},
\end{align*}
\]

was used for posterior and predictive analysis of subsets of this full data set. Here Dir denotes a Dirichlet distribution and IW an inverse Wishart, in standard notation.

The standard Gibbs sampler for mixture models (Lavine and West, 1992; West, 1992) delivers Monte Carlo approximations to posterior and predictive distributions. In particular, given posterior samples of all model parameters (including the latent mixture component indicators for each sample), the posterior mean of the regression curve can be approximated pointwise over a range of values \( x_* = [a ; b] \) to deliver the estimated regression function \( E(y_* | x_*, D) \) over this range. This is plotted in Figure 1(a) for the case in which \( D \) is the full set of 175 observations and \( x_* = [-2 ; 2] \). The corresponding estimates of the predictive density functions \( p(y_* | x_*, D) \) are plotted for three different values of \( x_* = \{-1.5, 0, 2\} \) in Figure 1(b). These regression and density curves can be viewed as the “gold-standards” as they fully utilize all the available data.

Assume now that we can only measure \( y \) values for \( x \) in the range \([-1, 1]\). This leaves us with a smaller labelled data set and an unlabelled set \( X^u \). We can fit the model to the labelled data only, or to the labelled and unlabelled data. The Gibbs sampler can be easily extended to treat the \( y \) values for the unlabelled \( X^u \) as missing data and draw the corresponding labels. Such analysis result in Monte Carlo estimates of regression curves and predictive densities that can be compared to these from the full data analysis: Figure 2 presents results using only the labelled data, and Figure 3 presents results using both the labelled and the unlabelled data. Comparison of these graphs with Figure 1 strikingly illustrates the differences, and highlights the improvements in prediction that can be obtained by incorporating the unlabelled data.

### 3.2 Classification and discrimination with mixtures

A related and also methodologically central mixture modelling context is that of classification and discrimination, in which \( x \) arises from a mixture distribution and \( y \) indicates the mixture component (West, 1992). For example, in binary classification, for each of \( y = 0, 1 \) the model
is specified with \( Pr(y = 1) = \pi, (0 < \pi < 1) \), and

\[ (x|y) \sim f_y(x) \]

for some parametrized densities \( f_0(\cdot) \) and \( f_1(\cdot) \). In the common Gaussian mixture model, \( f_0 \) and \( f_1 \) are multivariate normal densities parameterized by different means and variance matrices (Lavine and West, 1992), \( f_y(x) = N(x|\mu_y, \Sigma_y) \) for each \( y = 0, 1 \). Define \( \psi = \{\mu_0, \mu_1, \Sigma_0, \Sigma_1\} \). Here the scientifically natural specification of the joint distribution is via the “retrospective” construction of equation (2), parametrized as \( p(x|y, \psi) \) and \( p(y|\pi) \).

Relating to the prospective factorization of equation (1), the implied marginal for \( x \) and conditional for \( y \) given \( x \) are easily deduced; the former is the implied mixture of the two normal distributions weighted by probabilities \( \pi \) and \( 1 - \pi \), and the latter simply the revised outcome “classification” probability for \( y = 1 \) at a point \( x \), computed by Bayes’s theorem. It is transparent that unlabelled data matter in this setting. Observations from the marginal distribution for \( x \) provide information about both the mixture weight \( \pi \) and the parameters \( \psi \) of the component normals. Prediction of a new \( y_* \) at a point \( x_* \) is performed by estimating (whether by formal Bayesian computations or otherwise) the classification probability \( Pr(y_* = 1|x_*) \) which is a complicated function of all parameters \( \pi \) and \( \psi \), and depends critically on various non-linear functions of \( \psi \) in particular. Hence information about \( (\pi, \psi) \) from unlabelled data, as from any other source, feeds through to impact on predictions.

To connect with the general notation and theory of Section 2, we see that for the primary parameters \( \pi, \psi \) there is no simple reduction or separation of the parameters into distinct parameters \( \phi \) and \( \theta \). Each of the distributions \( p(y|x, \phi) \) and \( p(x|\theta) \) depend in complicated ways on all the parameters \( \pi \) and \( \psi \), and consistency with the notation in Section 2 is achieved only by setting \( \phi \equiv \theta = \{\pi, \psi\} \). Hence, in this key example, \( \phi \) and \( \theta \) are fundamentally highly structurally related, and the general theory of Section 2 implies that predictions will be impacted by the use of unlabelled data \( X^m \) in concordance with our immediate, context-specific deductions above.

We note that various statistical and algorithmic approaches have been proposed to take advantage of the information in \( X^m \) and the effectiveness of \( X^m \) has been either implicit or explicitly well-studied methodologically (M"ueller et al., 1996; Nigam et al., 2000; O’Neill, 1978; Ganesalingam and McLachlan, 1978, 1979). An interesting theoretical connection in Castelli and Cover (1995) concerns the asymptotics of prediction errors for \( y_* \) with respect to an increasingly large unlabelled sample, so that asymptotically all parameters are effectively “known”; the basic conclusion of this analysis was that labeled samples are exponentially more valuable than unlabeled samples in classification problems (Castelli and Cover, 1995).
3.3 Normal linear regression models

In the usual normal linear regression model, \( \phi = (\beta, \tau) \) is the set of regression parameters from the model

\[
y|x, \phi \sim N(\beta'x, \tau^2)
\]

where \( x \) and \( \beta \) are \( k \)-dimensional vectors. One way such a model can arise is from an assumed joint multivariate normal distributions for \((y, x)\), namely the \((k+1)\)-dimensional (zero-mean) normal \( N(0, \Sigma) \) where

\[
\Sigma = \begin{pmatrix}
\sigma_y^2 & \rho' \\
\rho & \Sigma_x
\end{pmatrix},
\]

for some scalar parameter \( \sigma_y \), \( k \)-dimensional vector of covariance parameters \( \rho \), and \( k \times k \) variance matrix \( \Sigma_x \). Under such a model we have \( \beta = \Sigma_x^{-1} \rho \) and \( \tau^2 = \sigma_y^2 - \beta' \rho \), and the marginal \( p(x|\theta) = N(\mu_x, \Sigma_x) \) with the characterizing parameter \( \theta = (\mu_x, \Sigma_x) \).

Some example contexts are as follows:

- A direct specification of the prior \( p(\phi, \theta) = p(\phi)p(\theta) \) that assumes independence, and so implies that unlabelled \( X^m \) data will be irrelevant to prediction of future \( y_* \). This would be typical in many applied regression settings.
- An indirect specification in which the initial prior is defined for \((\mu, \Sigma)\), with the prior \( p(\phi, \theta) \) being implied by transformation. A common approach is to use the conjugate normal-inverse Wishart prior distribution. Any prior in this class has the property that the implied prior on \((\phi, \theta)\) is in fact one in which \( \phi \) and \( \theta \) are independent (Geiger and Heckerman, 2002; Dobra et al., 2004).
- Other indirect specifications of the prior \( p(\phi, \theta) \) by deduction from a prior on \( \Sigma \) will induce dependence between \( \phi \) and \( \theta \) and hence lead to relevance of the unlabelled data since \( X^m \) will then provide information about \( \phi \) indirectly through its relevance for \( \theta \).

The second example here illustrates a case in which modelling prior information on parameters of the joint distribution of \( y \) and \( x \) using a standard conjugate implies that the unlabelled \( X^m \) data will be irrelevant for predicting \( y_* \). This result arises more generally in exponential family models. Other priors may, and usually will, lead to prior and therefore posterior dependence so the unlabelled data will be relevant.

3.4 Binary outcomes: Cancer incidence and prognosis

An illuminating example is the case of binary \( y \) and binary \( x \). For thematic context, suppose \( x = 1/0 \) represents the presence/absence of mutation in the BRCA1 breast cancer gene in a
woman, and that \( y = 1/0 \) represents occurrence of breast cancer before age 70. The goal here is to predict the probability of \( y = 1 \) given the presence or absence of the mutation.

In this breast cancer example we define parameters \( \theta \) as the incidence rate of the BRCA1 mutation, \( \phi_0 \) is the base rate for breast cancer in the general (wild type) population of women, and \( \phi_1 \) the (higher) cancer rate among carriers of the mutation. The joint density derived from the prospective construction using \( p(y|x, \phi) \) and \( p(x|\theta) \), is then parameterized by the three probabilities, \( \phi = (\phi_0, \phi_1) \) and \( \theta \) where

- \( \phi_x = p(y = 1|x, \phi) \) for \( x = \{0, 1\} \), and
- \( \theta = p(x = 1|\theta) \).

For this model, the predictive distribution given the data is

\[
p_* = p(y_* = 1|x_*, D) = \int\int \phi_{x_*} p(\phi, \theta|x_*, D) d\phi d\theta.
\]

Given the above model the following are two natural prior specifications:

- We can directly specify independent priors on \( \theta \) and \( \phi \). As a result \( p_* \) will not depend on the unlabelled data.
- We have cell probabilities \( p(x, y) \) on the joint space \( x = 0, 1; y = 0, 1 \) defined as

\[
\pi = \{\pi_{0,0}, \pi_{0,1}, \pi_{1,0}, \pi_{1,1}\}.
\]

Common approaches utilize Dirichlet priors on \( \pi \). If we choose a Dirichlet prior \( p(\pi) \) and find the implied prior \( p(\phi, \theta) \) by transformation, the result is prior independence of \( \phi \) and \( \theta \), and again the unlabelled data is irrelevant to prediction.

A more interesting and perhaps natural modelling assumption on the joint space is that the breast cancer samples come from an inhomogeneous population having two genetically and environmentally different sub-populations in connection with inherited breast cancer related characteristics and lifetime cancer risks. In this case a reasonable prior would be a mixture of two Dirichlets,

\[
p(\pi) = ap_0(\pi) + (1 - a)p_1(\pi)
\]

where \( p_0 \) and \( p_1 \) are two different Dirichlet priors for the two sub-populations, though the sampling design cannot distinguish between the sub-populations. It then follows by transformation that

\[
p(\phi|\theta) = w(\theta)p_0(\phi) + (1 - w(\theta))p_1(\phi)
\]

where \( p_0 \) and \( p_1 \) are the implied margins on \( \phi \) from each of the two Dirichlets, and the mixing probability \( w(\theta) \) is computed conditionally on any value of \( \theta \) as

\[
\frac{w(\theta)}{1 - w(\theta)} = \frac{a}{(1 - a) \left(1 - p_1(\theta)\right)}.
\]
Thus under a mixture prior of this form, $\theta$ and $\phi$ are dependent and so the unlabelled data $X^m$ will provide information about $y_\ast$ indirectly via $\theta$ and $\phi$. The dependence between $\phi$ and $\theta$ is reflected in the variation of the weight $w(\theta)$ that provides the “link” for the unlabelled data information to flow through to impact on inferences about $\phi$, and hence to $y_\ast$. Only in the extreme case of no subpopulation structure, when $p_0(\cdot) = p_1(\cdot)$, will the unlabelled data on the mutational incidence rate play no role in predicting cancer events for future patient.

4 Factor models and factor regression

4.1 Statistical framework

The interest in factor regression has increased due to the prevalence of problems with high-dimensional predictors. One common example is principal component regression (PCR). In PCR, the singular value decomposition of the design matrix of original predictor variables generates principal components – or empirical factors – that become the predictors in a regression. The resulting orthogonal regression and potential data reduction are two key benefits of this modelling approach. However, a key question is raised in connection with prediction: since we aim to predict $y_\ast$ values at new, future $x_\ast$ values, should we not include the future design points in the initial analysis and principal component evaluation? This is evidently just a question of whether, and if so, how, to use unlabelled data in the model development and analysis of existing labelled data.

The question, and the general discussion of PCR and empirical factor regression, can be embedded in the broader theoretical context of (latent) factor regression models. West (2003) formalized the development of large-scale, latent factor models coupled with regression on latent factors, and delineated a comprehensive framework for predictive modelling that was particularly motivated by problems involving larger numbers of predictors - the “large $p$, small $n$” paradigm. This elucidated the theory underlying PCR and modelling using principal component projections of high-dimensional covariates/predictors as a limiting case of a broader class of regression models where the predictors are latent variables. This framework and theory also clarified and justified the use of so-called $g$-priors (Zellner, 1986) for Bayesian shrinkage regression, and defined novel classes of multiple shrinkage methods that are significantly beneficial in prediction problems through the ability to induce differential shrinkage in different factor-predictor dimensions. Importantly, the framework trivially clarifies the issue of use of unlabelled data, and how unlabelled samples enter into predictions based on analysis of labelled data, in general. The special limiting case of principal component regression is one important benefit.

The following normal linear model serves as a specific example to illustrate the more gen-
eral principles of factor regression models. A univariate response \( y \) is to be predicted based on a (high-dimensional) \( p \times 1 \) predictor variable \( x \), and we have

\[
y_i = \alpha' \lambda_i + \epsilon_i \quad \text{and} \quad x_i = B \lambda_i + \nu_i
\]

where \( \epsilon_i \sim N(0, \sigma^2) \), \( \lambda_i \sim N(0, I) \) is a \( k \times 1 \) unit multivariate normal latent factor for each \( i \), \( B \) is an uncertain \( p \times k \) matrix of factor loadings of \( x \) on \( \lambda \), \( \nu_i \sim N(0, \Psi) \) is a vector of idiosyncratic noise terms, and \( \Psi \) is an uncertain diagonal variance matrix. Also, the \( \nu_i \) and \( \epsilon_i \) are conditionally (on all model parameters) mutually independent and independent over \( i \).

### 4.2 Unlabelled data in factor regression models

This framework is a key example of when unlabelled data matter. Fundamentally, the outcomes \( y \) to be predicted are modeled as responses in regressions on latent variables \( \lambda \), and the observed concomitant \( x \) variables are related to \( \lambda \), while \( y \) and \( x \) are conditionally independent given \( \lambda \). Thus the predictive relevance of \( x \) is indirect, through \( \lambda \).

By marginalizing over \( \lambda \) in the joint multivariate normal distribution of \( y, x \) and \( \lambda \) implied by the model specification, it becomes clear that we can identify \( p(y|x, \phi) \) as a normal linear regression of \( y \) on \( x \) with regression parameter vector and residual variance making up the prospective parameter \( \phi = \phi(\alpha, \sigma, B, \Psi) \). Also, the implied marginal distribution for \( x \) is normal with zero mean and variance matrix \( \theta = BB' + \Psi \). Thus, if \( \{B, \Psi\} \) are known then \( \theta \) is known and so the observed, unlabelled data \( X^m \) has no influence whatsoever in the problem of predicting a future \( y_* \) given data from either prospective or retrospective designs. However, typically \( \{B, \Psi\} \) are uncertain and need to be estimated. In this setting:

- Unlabelled data \( X^m \) provide information relevant to estimation of the latent factor model parameters \( \{B, \Psi\} \), and hence of relevance to predicting future \( y_* \) values via the transfer of information through inferences on the future \( \lambda_* \) related to \( x_* \).
- \( \phi \) is dependent of on aspects of \( \theta \) indirectly through their functional associations with the factor model parameters, so that any relevant prior \( p(B, \Psi, \alpha, \sigma) \) will induce dependencies between \( \phi \) and \( \theta \).

### 4.3 Digit classification example

The MNIST data set (Y. LeCun, [http://yann.lecun.com/exdb/mnist/](http://yann.lecun.com/exdb/mnist/)) is a standard data set used extensively in the machine learning community to benchmark binary regression models. The data set contains 60,000 images of handwritten digits \( \{0, 1, 2, ..., 9\} \), where each image consists of \( p = 28 \times 28 = 784 \) gray-scale pixel intensities. As an example, we consider what is generally regarded as one of the most difficult pairwise comparisons, that of discriminating a
handwritten “6” from a “9”. We frame this as a binary regression problem. The predictor space \( x \) is transformed via singular value decomposition of the initial design matrix of 784 primary pixel values (after centering), and the first two factors are used for predictive discrimination of unlabelled samples.

The data set contains 5918 handwritten “6”s and 5949 handwritten “9”s. Following Belkin et al. (2004), we take the first 400 observations from each class as a training sample and use the remaining samples as test cases to be predicted. The standard MCMC analysis of the probit regression model produces approximate posterior predictive probabilities of “6” versus “9” for each of the several thousand test samples, and we record empirical prediction error rates based on whether or not the predictive probability of the true digit (true label) lies below or above 0.5. For the labelled/unlabelled evaluation, our analysis is most extreme: we randomly select just two “6”s and two “9”s to treat as labelled, the remaining 398 in each of the two classes being regarded as unlabelled. To give an initial indication of the relevance of unlabelled data, Figure 4 plots the projections of the full sets of training and test data onto the two factors (first two principal components) of the labelled and unlabelled data together; the separation of digits is quite strong and clear for both the training and test data. Repeating the factorization and projection of the training data, but now using only 4 labelled samples (randomly selected with 2 of each digit), produces the graph in Figure 6(a); frames (b) (c) and (d) of Figure 6 show similar plots for different random draws of the 4 samples treated as labelled. The relevance of unlabelled samples is quite evident from comparison of these plots with those based on the labelled and unlabelled data together.

In the probit factor regression models using the first two principal components as predictors, and with a simple, standard normal/inverse gamma prior on regression parameters (West, 2003), repeated analysis of labelled data alone in the above framework yields an average prediction error rate on the test samples of approximately 31.2%. Repeating this analysis but now including the unlabelled data in defining the empirical factors yields a semi-supervised average error rate of approximately 9.5%. This gives some indication of the potential improvements in raw predictive accuracy that may accrue from the appropriate use of unlabelled data.

5 Kernel regression for prediction and classification

5.1 Kernel regression models

An interesting class of examples, which are central to the methodological interfaces of statistics and machine learning, arise in models based on kernel regression. Kernel and related smoothing spline methods have a long history in applied statistics and have seen a tremendous amount of development at the interfaces of machine learning and statistics in the last several
years (Poggio and Girosi, 1990; Wahba, 1990; Vapnik, 1998; Schölkopf and Smola, 2002; Shawe-Taylor and Cristianini, 2004; Liang et al., 2005).

The context is non-parametric, non-linear regression with \( y \in \mathbb{R}, x \in \mathbb{R}^k \), and a model of the form

\[
y = f(x) + \epsilon, \tag{6}
\]

where \( \epsilon \) is zero-mean noise term and \( f(\cdot) \) is an uncertain regression function. As an example, the class of Bayesian radial basis (RB) models (Liang et al., 2005) deals with questions of proper probability models - and the resulting proper inference and predictive results that then arise - for uncertain knots in a kernel model. This framework, and other approaches, begins with the interest in a representation of the form

\[
f(x) = \int w(u) K(x, u) dG(u), \tag{7}
\]

for some weight function \( w(u) \) over \( k \)-dimensional \( u \), and some specified kernel function \( K(\cdot, \cdot) \). The element \( G(\cdot) \) is a probability distribution function in \( k \)-dimensions. The key to the model is to note that, if \( G \) is discrete and put masses \( g_i \) at support points (or “knots”) \( u_i \), then the expression for \( f(\cdot) \) is simply

\[
f(x) = \sum_i g_i w(u_i) K(x, u_i),
\]

i.e., a radial basis function representation. The analysis of Liang et al. (2005) describes approximations to a model in which uncertainty about \( G \) is expressed using a Dirichlet process prior (Ferguson, 1973; Escobar and West, 1995). One implication of such a model for \( G \) is that, since Dirichlet processes are discrete with probability one, the formal mathematical model for \( f(x) \) is the sum above with a countably infinite number of knots \( u_i \). From the methodological viewpoint, both labelled and unlabelled \( x \) values provide information about \( G \) directly. In fact, with a sample of \( n \) labelled and/or unlabelled \( x \) values \( x_1, \ldots, x_n \) (whether from \( X, X^m \), or some combination of the two), this Dirichlet process model implies that \( f \) may be approximated by

\[
\hat{f}_n(x) = \sum_{i=1}^n w_{n,i} K(x, x_i), \tag{8}
\]

where \( w_{n,i} \propto w(x_i) \). The key methodological relevance of this approach is that this is true for all \( n \), providing consistency as sample size increases and additional design points are observed. This leads to the practical model in which each \( y_s \) is linearly regressed on the set of kernel predictors \( \{ K(x_s, x_i) \}_{i=1}^n \) based on whatever set of design points are observed. A complete model now involves a prior distribution over the induced regression coefficients \( w_{n,i} \) and we note that this explicitly depends on \( n \) and the realized \( x_i \). Hence, in both the structure of the regression model and in the requirements for a prior over coefficients, we see the dependence
on all values observed in the $x$ space; this is therefore a perfect example of when, why and how unlabelled $X^m$ data matter. In particular, we note that:

- $\theta = G(\cdot)$ so that $p(x|\theta)dx = dG(x)$ - the parameter is the full distribution function itself;
- Equations (6) and (7) show explicitly how $p(y|x, \phi)$ depends intimately on $\theta = G$ as defining the nonlinear kernel regression; in fact, $\theta \subseteq \phi$ in this case. Thus prior and posterior dependence of $\theta$ and $\phi$ is central to the model.
- As a result, unlabelled $X^m$ provides direct, immediate, and critically relevant information in predicting $y_*$.

This specific example of a kernel regression model derived within a coherent probabilistic framework here, taken from Liang et al. (2005), is presented for its simplicity and also as it represents a fully specified probabilistic model in which the kernel weights $w_{n,i}$ are related coherently as sample sizes change. Some additional connections and related kernel regression formulations are now mentioned.

### 5.2 Relation to machine learning kernel regression algorithms

Other constructions of kernel regression models, including those utilizing Gaussian processes and spline smoothing, non-Bayesian uses of radial basis functions and others (Poggio and Girosi, 1990; Wahba, 1990; Schölkopf and Smola, 2002; Shawe-Taylor and Cristianini, 2004; Vapnik, 1998), exhibit the same structure and consequent dependence on unlabelled data. One interesting connection with recent theoretical developments in machine learning approaches arises by noting that the central model of equation (7) also corresponds to the solution of the non-linear manifold regularization formulation of Belkin et al. (2004). This approach, motivated by geometric arguments, is an optimization algorithm that minimizes

$$f_* = \arg \min_{f \in \mathcal{H}_K} \left[ \frac{1}{n} \sum_{i=1}^{n} V(f(x_i), y_i) + \gamma_A ||f||^2_K + \gamma_I ||f||^2_I \right],$$

where $\{(y_i, x_i)\}_{i=1}^{n}$ are the labelled data, $\mathcal{H}_K$ is a Reproducing Kernel Hilbert Space (RKHS), $V(f(x), y)$ is a loss function, $||f||^2_K$ is the RKHS norm, $\gamma_A, \gamma_I$ are regularization parameters, and $||f||^2_I$ is a norm that reflects the smoothness of the function on the marginal $p(x)$. If the marginal is concentrated on a manifold, $x \subset M \in \mathbb{R}^k$, then a natural choice for $||f||^2_I$ is the Laplacian on the manifold. The marginal $p(x)$ is generally unknown; with unlabelled data $X^m$ from the marginal, the Laplacian on the manifold may be approximated by a Laplacian on the graph defined by the observed data (labelled and unlabelled)

$$\hat{f}_n(x) = \arg \min_{f \in \mathcal{H}_K} \left[ \frac{1}{n} \sum_{i=1}^{n} V(f(x_i), y_i) + \gamma_A ||f||^2_K + \frac{\gamma_I}{(n + n_m)^2} f^T L f \right].$$
where $L$ is the graph Laplacian on all the data (given a weight matrix on the graph) and $f = \{ f(x_1), ..., f(x_n), f(x_1^m), ..., f(x_{nm}^m) \}$. The above optimization is achieved by

$$
\hat{f}(x) = \sum_{i=1}^{n} w_{n,i}K(x, x_i) + \sum_{i=1}^{nm} w_{n+n_m, n+i}K(x, x_i^m),
$$

which takes the same form as the equation (8). This formulation as an optimization problem from a statistical machine learning viewpoint generates precisely the same functional form of the model as that derived from a non-parametric regression in the Bayesian framework above, and the consequences for the use of unlabelled data in model formulation are the same.

5.3 Illuminating the potential impact of unlabelled data

A simple but illuminating synthetic example provides an initial illustration. A data set of 50 points $\{(x_i, y_i), i = 1 : 50\}$ is plotted in Figure 6(b); here $x_i \in \mathbb{R}^2$ and $y_i = 0/1$. This data set can be easily classified according to $y = 0$ versus $y = 1$ by a Gaussian kernel model, and we fit such a model using the Bayesian model completion – in terms of prior specification for the kernel weights and observational variance parameters – and the resulting MCMC method for model fitting as described in Liang et al. (2005). Though the details of the prior specification and computation are not central here, we note that the model involves use of a generalized shrinkage prior, termed *generalized g-prior* by West (2003), on the kernel regression coefficients. This is a method of importance when dealing with large numbers of regression parameters, and its use in these kernel models where the number of regression parameters exceeds the number of labelled observations is particularly apt.

The analysis leads to the computation of posterior predictive probabilities for $y = 1$ versus $y = 0$ at any chosen new $x$ value, i.e., the class predictions based on any data set. Using the fully labelled 50 data points for such an analysis yields results displayed and described in Figure 6(a).

The analysis is repeated using only 4 labelled points, 2 each with $y = 0$ and $y = 1$; the 4 randomly selected points are marked in Figure 7(b). The resulting class prediction contours and summaries of predictions for the 46 unlabelled points are then computed in two separate analysis: (i) using only the labeled data – just the 4 points; and (ii) using the labelled and unlabelled data. Figures 7 and 8 present the results of these two analysis. This exercise was repeated using a total of 8 labelled points, resulting in the displays in Figures 9 and 10. From the figures the major impact of unlabelled data is clearly apparent, and its relevance with very small numbers of labelled samples highlighted. We also see that the semi-supervised analysis using only 8 labelled samples results in predictions that are very similar to those obtained if all 50 samples were labelled. Unlabelled data can dramatically impact upon and improve prediction accuracy.
5.4 Kernel regression for cancer classification using genomic data

A substantive example involves analysis of a gene expression data set consisting of DNA microarray expression profiles from 190 tissue samples representing a variety of different primary tumors (breast, prostate, lung, lymphoma, etc.) and 90 non-cancerous, “normal” samples from the corresponding tissue of origin (Ramaswamy et al., 2001; Mukherjee et al., 2003). Following standard processes of data normalization and screening for genes showing non-trivial variation, the data analyzed consists of \( p = 2800 \) gene expression variables, or “genes”, on the set of 280 samples. The analysis set up aims to use the gene expression data as predictors \((x)\) in a binary kernel regression model with outcome \( y = 1 \) representing “cancer”, i.e., any of the cancer types, and \( y = 0 \) representing normals. As in the synthetic example above, the analysis uses a Gaussian kernel model fitted using Bayesian shrinkage priors, as in Liang et al. (2005). Our interest is to compare predictions of cancer versus normal under this given model and prior specification applied to differing selections of data – selections that allow us to examine the impact of unlabelled data.

We do this as in the synthetic example above – randomly selecting a fraction of the data to be regarded as unlabelled, fitting the model and then predicting the status (cancer versus normal) of the selected unlabelled cases in terms of posterior predictive probabilities. We repeat this analysis twice – first, using only the labelled data; second, using both the labelled and unlabelled data – and are then able to compare predictions between the two analysis to assess the changes due to use of the unlabelled data. For a given fraction of unlabelled data, we repeated this 50 times, each time randomly selecting the cases to be labelled/unlabelled, and computing the average (across the 50 repeats) empirical prediction error rate in classifying the unlabelled cases. The prediction of an unlabelled sample is regarded as “correct” if the predictive probability of the true state (cancer or normal) exceeds 0.5. Figure 11 summarizes the resulting empirical error rates for a series of such analysis in which we progressively increased the percentage of labelled data from 10% to 90%. The figure clearly shows the differences between analysis using only labelled data and that using labelled and unlabelled.

Additional insight into the impact of including unlabelled samples is given in Figure 12. From one analysis with 80% of the data unlabelled, we selected 10 each of the cancer and normal samples that were unlabelled in the analysis, and graph the estimated predictive probabilities of cancer versus normal with approximate 95% credible interval. This shows the impact of the unlabelled \( x \) data on the predictions, in terms of the impact on estimates of prediction uncertainty as well as empirical accuracy.
6 Summary comments

Beginning with an articulation of the basic sampling and design specifications underlying statistical formulations of prediction problems, we have delineated the conceptual and theoretical issues underlying the use and relevance, or irrelevance, of unlabelled data in classification and prediction problems. This, coupled with a series of examples in central statistical modelling contexts, and empirical illustrations and evaluations in two substantive data analysis, provides an overview and synthesis of the ideas underlying the emerging methodology of semi-supervised learning in the machine learning and statistics literatures.

Graphical model representations of the joint sampling model context aid in this interpretation. The relevance, or otherwise, of the unlabelled $X^m$ data can be deduced essentially by inspection of the implied (undirected) graphical representation of any full model structure. For example, the full distribution assuming joint sampling, and in cases for which $p(\phi, \theta) = p(\phi)p(\theta)$, is illustrated in graphical terms in Figure 13. The joint density exhibited here is

$$p(y_s, x_s, Y, X, X^m, \phi, \theta) = p(y_s|x_s, \phi)p(Y|X, \phi)p(X|\theta)p(x_s|\theta)p(X^m|\theta)p(\phi)p(\theta).$$

Figure 13(a) is a directed acyclic graph of the joint distribution structured in terms of composition of sampling distributions. Figure 13(b) displays the corresponding undirected graph in which the lack of an edge between $X^m$ and $y_s$ indicates conditional independence given all other quantities, hence the irrelevance to prediction of the unlabelled data in this case. In contrast, were $\phi, \theta$ to be a priori dependent, then the five nodes of the undirected graph would be fully connected, exhibiting the relevance of the unlabelled data to prediction of $y_s$.

In addition to clarifying and exemplifying the structure of models and the prediction problem with unlabelled data, one aim of this work has been to review the area to provide a link across the mainstream statistical to machine learning communities. We hope that this will entice more statistical researchers into a very active, productive and exciting research milieu, while also founding the discussion in venerable, simple and unambiguous terms arising from the direct and classical probabilistic formulation. This view directly, we believe, addresses and answers the questions of “when, why and how” unlabelled data help in predictive modelling.

Acknowledgments

The authors acknowledge the most constructive suggestions of the Executive Editor and an anonymous Editor on the original version of this paper. The research reported here was partially supported by National Science Foundation (grants DMS-0342172 and DMS 0406115).
Any opinions, findings and conclusions or recommendations expressed in this work are those of the authors and do not necessarily reflect the views of the NSF.

References


Figure 1: Mixture model regression and prediction example - analysis using the full data set: (a) Scatter plot of all 175 data points together with the posterior predictive regression curve of $(y|x)$ and three contours representing the posterior estimates of the three Gaussian components. (b) The estimated predictive density functions $p(y|x)$ evaluated at three chosen values $x = -1.5, 0, \text{ and } 2.$
Figure 2: Mixture model regression and prediction example - analysis using only the labelled data:
(a) Scatter plot of the labelled data points together with the estimated regression curve and contours of the three estimated Gaussian components. (b) The estimated predictive density function of $p(y|x)$ evaluated at $x = -1.5, 0, \text{ and } 2$. 
Figure 3: Mixture model regression and prediction example - analysis using the labelled data and also the unlabelled points (red circles): (a) Scatter plot of the labelled data points together with the estimated regression curve and contours of the three estimated Gaussian components. (b) The estimated predictive density function of $p(y|x)$ evaluated at $x = -1.5, 0, \text{ and } 2$. 
Figure 4: MNIST handwritten digit data example, with samples of handwritten “6” (red) and “9” (blue). The axes are the first two principal components computed from a singular value decomposition of the centered data, using the full data set of over 11,000 samples (5918 “6”s and 5949 “9”s). Scatter plotted on these empirical factors are (a) the training data of 800 samples (400 of each digit), and (b) the remaining test data. The two factors evidently carry strongly discriminating information.
Figure 5: MNIST handwritten digit data example, in a format similar to that of Figure 4. In these frames the two principal component were evaluated on only 4 samples selected as labelled, 2 of each digit, and the scatter plots are of the training data of 800 samples. The 4 labelled samples were randomly drawn from the training data set, and the four frames here represent four different draws of the labelled samples. Though there is evidence of discriminatory information to distinguish the handwritten “6”s (red) from the “9”s (blue), it is quite clear that discriminatory power will be very limited.
Figure 6: Kernel regression example using synthetic data. Frame (b) displays a scatter plot of the 50 observations on the $x_2$ versus $x_1$ axes, with cases $y_i = 0$ in blue and $y_i = 1$ in red. The binary kernel regression model analysis of the full set of 50 labelled observations produces approximate posterior predictive probabilities $\Pr(y_s = 1|x_s, D)$ at any point $x_s$ in the plane; the green points in frame (b) are points at which $\Pr(y_s = 1|x_s, D) = 0.5$, i.e., represent points on the separating contour. Frame (a) displays a color image of the contours of $\Pr(y_s = 1|x_s, D)$ as $x_s$ varies, coded such that the predictive probability of $y_s = 1$ increases from near 0 (blue) to near 1 (red).

Figure 7: Kernel regression example with displays as in Figure 6, but now with the model fitted to only the four selected data points that are indicated with larger symbols – two red (1) and two blue (0) – in frame (b). The predictions in this case are then just based on these four samples.
Figure 8: Kernel regression example with displays as in Figure 7, again using only the same four selected data points as labelled data, but now also including the unlabelled $X^m$ data. Evidently the inclusion of $X^m$ has a substantial impact on the predictive probability contours and represents an improvement over the use of the labelled data alone.

Figure 9: Kernel regression example using only labelled data, but now using 8 labelled observations indicated in frame (b).
Figure 10: Kernel regression example using the 8 labelled data points together with the remaining unlabelled samples. Notice that the use of the 42 unlabelled samples is sufficient to produce predictive probability contours that are very similar to those using the full set of labelled observations, as in Figure 6 though with evident and justifiably greater uncertainty, even though the \( y \) values are labelled on only 8 data points.
Figure 11: Analysis and prediction of cancer versus normal tissue \((y = 0, 1)\) based on gene expression data \((x)\). Empirical prediction error rates from the analysis ignoring the unlabelled data appear in red, and the corresponding figures from the analysis that includes the unlabelled data in blue. Repeat analysis were performed with varying percentages of cases randomly selected to be designated as labelled \((x\)-axis) with the remainder being unlabelled. For each such percentage, the selection of unlabelled cases and the Bayesian MCMC analysis of the kernel regression model used for prediction was rerun 50 times, and the figures graphed represent the average prediction error in predicting the true status \((y = 1\) versus \(y = 0\) with predictive probability of at least 0.5 being regarded as “correct”) for the unlabelled samples. The uniform improvement in empirical predictive accuracy when including the unlabelled data is clear.
Figure 12: Cancer versus normal predictions using kernel model. The figure displays estimated predictive probabilities of cancer versus normal for 10 cancers (red) and 10 normal tissues (blue) that were unlabelled in the data analysis. This analysis involved only 20% of the data being labelled. The frames also provide estimated 95% credible intervals associated with each of the predictions. This shows the impact of the unlabelled x data on the predictions, in terms of the impact on estimates of prediction uncertainty as well as empirical accuracy.
Figure 13: Graphical models of the basic structure relevant to understanding the role of unlabelled data in predictive modelling. The figure shows the directed (acyclic) graph (a) and undirected graph (b) of the joint distribution of data and parameters in cases of independence of $\phi, \theta$. In contrast, if $\phi, \theta$ are dependent, then (a) would have an edge between $\phi$ and $\theta$, and (b) would be a fully connected graph.