Optimising Kernel Parameters and Regularisation Coefficients for Non-linear Discriminant Analysis

Tonatiuh Peña Centeno and Neil D. Lawrence

Department of Computer Science
The University of Sheffield
Regent Court, 211 Portobello Street
Sheffield, S14 DP, U.K.
{tpena,neil}@dcs.shef.ac.uk

Abstract In this paper we consider a Bayesian interpretation of Fisher's discriminant. By relating Rayleigh's coefficient to a likelihood function and through the choice of a suitable prior we use Bayes' rule to infer a posterior distribution over projections. Through the use of a Gaussian process prior we show the equivalence of our model to a regularised kernel Fisher's discriminant. A key advantage of our approach is the facility to determine kernel parameters and the regularisation coefficient through optimisation of the marginalised likelihood of the data.

1 Introduction

Data analysis typically requires a preprocessing stage to give a more parsimonious representation of data, such preprocessing consists of selecting a group of characteristic features according to an an optimality criterion. Tasks such as data description or discrimination commonly rely on this preprocessing stage. For example, Principal Component Analysis (PCA) describes data more efficiently by projecting data onto the principal components and then by minimising the reconstruction error, [4]. In contrast, Fisher's discriminant separates classes of data by selecting the features that maximise the ratio of projected class means to projected intraclass variances, [1].

The intuition behind Fisher's Linear discriminant (FLD) consists of looking for a direction of discrimination $\mathbf{w}$ such that, when a set of training samples are projected on to it, the class centres are far apart while the spread within each class is small consequently producing a small overlap between classes [16]. This is done by maximising a cost function known as Rayleigh's coefficient, $J(\mathbf{w})$. Kernel Fisher's discriminant (KFD) is a nonlinearisation that follows the same principle but in a possibly high-dimensional feature space $\mathcal{F}$. In this case, the algorithm is reformulated in terms of $J(\boldsymbol{\alpha})$, where $\boldsymbol{\alpha} \in \mathcal{F}$ is the new direction of discrimination. The theory of reproducing kernels in a Hilbert space gives the relation between both vectors $\mathbf{w}$ and $\boldsymbol{\alpha}$, [16]. In either case, the objective is to determine the most 'plausible' direction according to the optimality criterion, J.

Kernel Fisher's discriminant has been applied successfully to classification problems [11] and more recently, due to interpretation of its structure, it has been formulated as an optimisation problem that leads to a special form of regression [12]. KFD shares many of the virtues of other kernel based algorithms: an appealing interpretation of a kernel as a mapping of an input to a high dimensional space and a good performance in real life applications, among the most important. However, KFD also suffers from some of the deficiencies of kernelised algorithms: the solution will typically include a regularisation coefficient to limit model complexity and parameter estimation will rely on some form of cross validation. The former introduces an extra parameter that must be estimated while the latter makes parameter specification a lengthy process.

In this paper we introduce a novel probabilistic interpretation of Fisher's discriminant. The probabilistic model is outlined in Section 2 along with a brief review of the existing literature on FLD. We build up over this model in Section 3 by first applying priors over the direction of discrimination to develop a *Bayesian* Fisher discriminant. In later sections, we show that the introduction of a Gaussian Process prior renders the model equivalent to a regularised KFD. Section 4 details an algorithm for estimating the parameters of the model (kernel and regularisation coefficient) by optimising the marginal log likelihood. We present the results of our approach by applying it on toy data and by classifying standard datasets from several repositories in Section 5. Finally we address future directions of our work.

2 Probabilistic Interpretation

Given a set of training data $(\mathbf{X}, \mathbf{t}) = \left\{ \left(\mathbf{x}^{(i)}, t_i \right) \middle| \, i = 1 \ldots N \right\}$, standard linear regression consists on specifying a vector of parameters $\mathbf{w}$ that best explains the relation $t_i = f\left(\mathbf{x}^{(i)}; \mathbf{w}\right) + \epsilon$. The noise ϵ is commonly assumed to be additive Gaussian noise $N\left(0, \beta^{-1}\right)$ [1] with precision β and the function is commonly specified as $f_i = \mathbf{w}^{\mathrm{T}} \mathbf{x}^{(i)}$. Two common methods to estimate the vector of weights are Maximum Likelihood (ML) and Bayesian inference. Maximum Likelihood approaches the problem from an optimisation perspective and yields a point estimate $\hat{\mathbf{w}}$, whereas a Bayesian approach gives the probability distribution over the $\mathbf{w}$. Standard regression can be converted into a classification task if all the targets t_i are encoded as categorical variables, e.g. $t_i \in \{0, 1\}$.

In this section, we propose a likelihood function closely related to the linear regression model just described but with noise arising from a mixture of two Gaussian distributions with equal variance. We will argue that maximisation of this likelihood renders an estimate equivalent up to a constant of proportionality to that produced from maximising Rayleigh's coefficient. We will use this probabilistic model as a basis for later sections.

[1] We use the notation $N(\mathbf{x}|\mathbf{m}, \Sigma)$ to indicate a multivariate Gaussian distribution over $\mathbf{x}$ with mean $\mathbf{m}$ and covariance Σ.

2.1 Fisher's discriminant Analysis

Fisher's discriminant analysis involves seeking a direction or feature $\mathbf{w} \in \mathbb{R}^{d\times 1}$ for which separation in a data set will be maximised. The discriminant looks to provide a good separation between the projected class means while achieving a small variance around those projections. The hope is that it will be possible to distinguish the different classes from these projections with small error. We first introduce some notation that will help us describe this idea in a mathematical way. Let a set of training samples $(\mathbf{X}, \mathbf{y}) = \left\{ \mathbf{x}^{(i)}, y^{(i)} \,\middle|\, i = 1 \dots N \right\}$ be split into two groups $\mathbf{X}_q = \left\{ \mathbf{x}_q^{(i)} \,\middle|\, i = 1 \dots N_q \right\}$ for $q \in \{0, 1\}$ such that $\mathbf{X} = \mathbf{X}_0 \cup \mathbf{X}_1$. Furthermore, consider the sample mean for each class to be $\mathbf{m}_q = N_q^{-1} \sum_{i \in N_q} \mathbf{x}_q^{(i)}$ and the class label vectors $\mathbf{y}_1 = \mathbf{y}$ and $\mathbf{y}_0 = \mathbf{1} - \mathbf{y}$, where $\mathbf{y} \in \{0, 1\}^N$ and $\mathbf{1}$ is a vector filled with ones. Then by considering a linear function $f = \mathbf{w}^{\mathrm{T}}\mathbf{x}$ that projects all data onto the direction of discrimination, Rayleigh's criterion can be written as the ratio

$$J = \frac{(\mu_1 - \mu_0)^2}{\sigma_1^2 + \sigma_0^2}, \tag{1}$$

where the parameters $\mu_q = N_q^{-1}\mathbf{w}^{\mathrm{T}}\mathbf{m}_q$ and $\sigma_q^2 = \sum_{i \in N_q} \left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_q^{(i)} - \mu_q \right)^2$ are the mean and variance of the projected data [3, 1]. Making explicit the dependence of J on $\mathbf{w}$ gives

$$J(\mathbf{w}) = \frac{\mathbf{w}^T \Sigma_B \mathbf{w}}{\mathbf{w}^T \Sigma_w \mathbf{w}}, \tag{2}$$

with

$$\Sigma_B = (\mathbf{m}_1 - \mathbf{m}_0)(\mathbf{m}_1 - \mathbf{m}_0)^{\mathrm{T}} \quad \text{and}$$

$$\Sigma_w = \sum_{q \in \{0,1\}} \sum_{i=1}^{N} \left(\mathbf{x}_q^{(i)} - \mathbf{m}_q \right) \left(\mathbf{x}_q^{(i)} - \mathbf{m}_q \right)^{\mathrm{T}}.$$

The between covariance matrix Σ_B measures separation between projected means and the intraclass covariance Σ_w gives an estimation of the separation around those projections. The formulation in (2) makes evident that this problem uniquely involves the vector $\mathbf{w}$. Moreover it shows that a solution can be found by solving a generalised eigenproblem of the form $\Sigma_B \mathbf{w} = \lambda \Sigma_w \mathbf{w}$, with λ being the eigenvalues of $\Sigma_w^{-1}\Sigma_B$. An alternative solution is given by $\mathbf{w} \propto \Sigma_w^{-1}(\mathbf{m}_0 - \mathbf{m}_1)$.

A more detailed analysis of Fisher's discriminant allows it to be cast as a quadratic programming problem. We notice first that Σ_B has rank one, such that the numerator of (2) can be held constant. As in a discrimination task the magnitude of the solution $\mathbf{w}$ is not relevant, fixing the numerator to an arbitrary scalar d and minimising $\mathbf{w}^T \Sigma_w \mathbf{w}$ will yield an equivalent result. Based on these observations, Mika [12, 10] has shown that Fisher's discriminant can be cast in a more general framework by considering it a convex, quadratic optimisation problem. In this case, the variance of the projections is minimised while the

distance between projected means is kept at an arbitrary value, say $\mu_0 - \mu_1 = d$. In other words, the coefficient becomes $J = d^2 / \sigma_1^2 + \sigma_0^2$. We will also make use of this 'average distance' constraint in subsequent sections to reformulate Fisher's discriminant.

2.2 A Likelihood Model

Several sources in the literature have drawn relationships between FLD analysis and least squares regression. It is interesting to see the similarities between minimising the class separation in the output space, just as in FLD and setting the outputs as close as possible to some target, as in least squares. More specifically, [1, 3] have showed that FLD arises as a special case of regression in which targets have been encoded in a particular way. From this formulation it has also been possible to draw connections between Fisher's discriminant and the so called least-squares Support Vector Machine (ls-SVM) [17], with the latter being an SVM that implements a particular loss function.

In this section, we encode the targets in a similar way to that of [3, 1] in order to approach FLD from a probabilistic point of view. We first introduce a likelihood function that explains why some projected data $\mathbf{f} = \{ f_i | \, i = 1, \ldots, N\}$ is clustered around the two class centers c_q. In order to do so, we use two Gaussian distributions with unique precision that are centered at c_q such that each one of them measures the amount of separation between the projections and the class centers. Despite not being a very appealing model at first sight, we will demonstrate that optimising the locations c_q will give similar results to that of Fisher's discriminant, which is based on the rather complementary view of optimising the location of each data projection f_i.

Let the target vectors be encoded as $\mathbf{t}_q = c_q \mathbf{y}_q$ and also let the precision of each Gaussian be β. Then the distribution over the targets can be expressed by

$$p\left(\mathbf{t}|\mathbf{f}, \beta^{-1}\right) = \frac{\beta^{\frac{N}{2}}}{(2\pi)^{\frac{N}{2}}} \exp\left\{ -\frac{\beta}{2} \sum_{q \in \{0,1\}} \left(\mathbf{t}_q - \mathbf{f}\right)^{\mathrm{T}} \mathrm{diag}\left(\mathbf{y}_q\right) \left(\mathbf{t}_q - \mathbf{f}\right) \right\}. \quad (3)$$

We know that this is not a probability distribution in strict sense as it is not normalisable, i.e. $\mathbf{t}$ is a discrete variable. Nonetheless, we 'relax' this assumption and take it to be a proper distribution in order to be able to build upon our model. We will see that sensible results can be obtained despite this relaxation in Section 5. Moreover, in future work, due to this assumption we think it will be possible to construct a noise model over the targets that is similar to [6].

Equation (3) can be re-written in terms of the parameters of the model, c_0, c_1 and β (Appendix A)

$$L\left(c_0, c_1, \beta|\mathbf{f}, \mathbf{y}\right) = \frac{\beta^{\frac{N}{2}}}{(2\pi)^{\frac{N}{2}}} \exp\left\{ -\frac{\beta}{2} \left[\sum_{n=1}^{N} y_n \left(c_1 - f_n\right)^2 + \sum_{n=1}^{N} \left(1 - y_n\right) \left(c_0 - f_n\right)^2 \right] \right\}. \quad (4)$$

In a typical Maximum likelihood setting, we look to maximise L in order to obtain the most probable set of parameters. Differentiating the log of (4) with respect to c_q and equating the corresponding gradient to zero gives

$$\hat{c}_q = \frac{1}{N_q}\mathbf{y}_q^{\mathrm{T}}\mathbf{f} \equiv \mu_q, \tag{5}$$

which is the sample mean of each class. In the remaining parts of this paper we will be using this definition of $\hat{c}_q$. A solution for the parameter β is found by applying the same procedure, such that

$$\hat{\beta} = \frac{N}{\sum_{n=1}^{N} y_n \left(c_1 - f_n\right)^2 + \sum_{n=1}^{N} \left(1 - y_n\right)\left(c_0 - f_n\right)^2}. \tag{6}$$

So far we have argued that a discriminant of some kind can be obtained by placing Gaussians with equal precisions around the projected data. The solutions obtained tell that the most probable centers are located at the sample means and that the noise level is inversely proportional to the separation of the class centers from each projection. However, it is still necessary to demonstrate that this procedure gives a solution identical to that of Fisher's discriminant mentioned at the end of section 2.1. In order to do so, we first back substitute the values $\hat{c}_q$ into (6) and recall the definition of σ_q^2, such that the resulting expression is

$$\hat{\beta}\left(\mathbf{f}\right) = \frac{N}{\sigma_1^2 + \sigma_0^2}. \tag{7}$$

We notice that equation (7) is proportional to the constrained definition of Fisher's discriminant. Therefore, under our proposed framework, Rayleigh's coefficient can be written exclusively in terms of the noise level $\hat{\beta}$, that is

$$J = \frac{d^2\hat{\beta}}{N}. \tag{8}$$

A solution to FLD can then be found not only by solving a generalised eigenproblem but also by adjusting the precision of the two Gaussians and by setting the locations of their centers. We can also express our proposed likelihood in terms of $\hat{\beta}$ by substituting the values of $\hat{c}_q$ and $\hat{\beta}$ into (4) , i.e.

$$L\left(\mathbf{f}\right) = \frac{\hat{\beta}^{N/2}}{2\pi^{N/2}} \exp\left\{-\frac{N}{2}\right\}.$$

As $L\left(\mathbf{f}\right)$ increases with $\hat{\beta}$ just as J does, we conclude that maximising the likelihood in (4) is equivalent to maximising Rayleigh's quotient.

2.3 Optimising w

In the previous section we showed that FLD analysis is equivalent to maximising a particular likelihood. Now, by substituting (5) into (4) and making $f_n =$

$\mathbf{w}^{\mathrm{T}}\mathbf{x}^{(n)}$we will express the likelihood in terms of a linear model. Maximisation of the resulting expression with respect to $\mathbf{w}$ will recover the standard solution for Fisher's discriminant. First we recall that the average projected class means must still lie apart at a distance $d = \mu_0 - \mu_1$. This condition is imposed as a constraint on the optimisation in the form of a Lagrange multiplier λ, hence the resulting constrained log-likelihood takes the form

$$\begin{aligned} \Lambda(\mathbf{w}, \beta, \lambda) = \frac{N}{2} \ln\left(\frac{\beta}{2\pi}\right) - \frac{\beta}{2}\left[\sum_{n=1}^{N} y_n \left(\mathbf{w}^{\mathrm{T}}\mathbf{x}^{(n)} - \mathbf{w}^{\mathrm{T}}\mathbf{m}_1\right)^2 \right. \\ \left. + \sum_{n=1}^{N} (1 - y_n) \left(\mathbf{w}^{\mathrm{T}}\mathbf{x}^{(n)} - \mathbf{w}^{\mathrm{T}}\mathbf{m}_0\right)^2\right] + \lambda\left[\mathbf{w}^{\mathrm{T}}(\mathbf{m}_0 - \mathbf{m}_1) - d\right]. \end{aligned} \tag{9}$$

The solution for $\mathbf{w}$ can then be found as

$$\hat{\mathbf{w}} \propto \Sigma_w^{-1}(\mathbf{m}_0 - \mathbf{m}_1), \tag{10}$$

where the equivalence with Fisher's discriminant is evident. Letting $\Delta\mathbf{m} = \mathbf{m}_0 - \mathbf{m}_1$, the constant of proportionality is given by $d\left(\Delta\mathbf{m}^{\mathrm{T}}\Sigma_w^{-1}\Delta\mathbf{m}\right)^{-1}$ which in turn is dependent on the selected value of d. This completes our discussion for a probabilistic interpretation for Fisher's discriminant. We have formulated FLD in terms of a likelihood function (2.2) that depends on the projected class centers and on some noise level β^{-1}. Furthermore, by exploiting the structure of FLD, we showed that maximisation of this model is equivalent to optimisation of Rayleigh's coefficient. Lastly, expressing the proposed likelihood in terms of a linear model and optimising it with respect to $\mathbf{w}$ gives an equivalent result to Fisher's discriminant. We can therefore complement the model and build on it by introducing priors over the direction of discrimination.

3 Bayesian Formulation

First let's reconsider the example of linear regression, i.e. modelling of a function $f(\mathbf{X}; \mathbf{w})$ by estimating the parameters $\mathbf{w}$. In the Bayesian approach, it is customary to assume the observations to have been corrupted by Gaussian noise: $\mathbf{x} \sim N\left(\mathbf{0}, \beta^{-1}\mathbf{I}\right)$ and to assume a Gaussian prior over the weights: $\mathbf{w} \sim N(\mathbf{0}, \mathbf{I})$, as well. Section 2.2 approached this problem by maximising a likelihood associated to FLD. This section will follow suit by expressing the likelihood in terms of $\mathbf{w}$ in order to infer its probability distribution. The distribution mean will have a value proportional to the standard solution of Fisher's discriminant. A probabilistic approach of this kind will offer some intrinsic advantages: the ability to compute variances of the projected points and the possibility to introduce Gaussian Process priors in a natural way. Nonetheless, in order to obtain a sensible solution, it will be necessary to incorporate the 'distance constraint' to the inference process.

3.1 Weight space formulation

So far we have found the most probable direction $\hat{\mathbf{w}}$ from a ML perspective, (10). Now what we seek is a distribution over projections which is obtained through combining our proposed likelihood with a prior distribution. Hence, the posterior probability over the weights is found by applying Bayes' rule:

$$p\left(\mathbf{w}|\,\mathbf{t},\beta,\mathbf{X},\mathbf{y}\right)=\frac{p\left(\mathbf{t}|\mathbf{w},\beta,\mathbf{X},\mathbf{y}\right)p\left(\mathbf{w}\right)}{p\left(\mathbf{t}|\beta,\mathbf{X},\mathbf{y}\right)}. \tag{11}$$

Assuming the distribution $p\left(\mathbf{t}|\,\mathbf{w},\beta,\mathbf{X}\right)$ to be given by (2.2) and the prior over the weights $\mathbf{w}\sim N\left(\mathbf{w}|\mathbf{0},\mathbf{A}^{-1}\right)$ will yield in a posterior of the form

$$\begin{aligned}p(\mathbf{w}|\,\mathbf{t},\beta,\mathbf{X})\propto\exp\Bigg\{&-\frac{\beta}{2}\sum_{n=1}^{N}\left(1-y_n\right)\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}^{(n)}-c_0\right)^2\\&-\frac{\beta}{2}\sum_{n=1}^{N}y_n\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}^{(n)}-c_1\right)^2-\frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{A}\mathbf{w}\Bigg\},\end{aligned} \tag{12}$$

where we recognise in this formulation the probabilistic analogy of the ls-SVM described in [17]. In contrast with the standard Support Vector Machine of [18], least-squares SVM's implement a squares error cost function in order to obtain a solution in the feature space. In addition, they have been linked to ridge regression classification for binary targets and to a regularised form of Fisher's discriminant analysis, the only difference between FLD and ls-SVM being the encoding of the targets. In our case, the posterior (12) is already implementing a version of Fisher's discriminant with a regularisation term arising naturally from the prior over $\mathbf{w}$.

In the subsequent discussion we will drop out the dependence of the posterior from the centers c_q by substituting each value for their most probable one, i.e. $\hat{c}_q=\mathbf{w}^{\mathrm{T}}\mathbf{m}_q$. With some mathematical manipulation we will be able to express the posterior solely in terms of the direction of discrimination,

$$p(\mathbf{w}|\,\mathbf{y},\mathbf{X})=N\left(\mathbf{w}|\mathbf{0},\mathbf{B}^{-1}\right), \tag{13}$$

with $\mathbf{B}=\beta\mathbf{X}^{\mathrm{T}}\mathbf{L}\mathbf{X}+\mathbf{A}$ and $\mathbf{L}=\mathbf{I}-\frac{1}{N_1}\mathbf{y}_1\mathbf{y}_1^T-\frac{1}{N_0}\mathbf{y}_0\mathbf{y}_0^T$.

As it was mentioned previously, so far the model constructed has not considered what we have called the average distance constraint. Avoiding its inclusion leads to a symmetry problem which causes the expected value of the posterior to be $\mathbf{0}$. This is better explained by considering that both solutions $\mathbf{w}$ and $-\mathbf{w}$ are valid and hence in average they 'nullify' each other. In a maximum likelihood scenario, as we have already shown, a constrained optimisation might take place by the inclusion of a Lagrange multiplier, just as it was showed before. However Bayesian models must deal with every parameter involved in terms of probability

distributions, hence we introduce the constraint through a function that places all of its mass in a single point, e.g.

$$p\left(d|\,\mathbf{w},\mathbf{m}_0,\mathbf{m}_1\right)=\lim_{\gamma\to\infty}\frac{\gamma^{\frac{1}{2}}}{\sqrt{2\pi}}\exp\left(-\frac{\gamma}{2}\left(d-\mathbf{w}^{\mathrm{T}}\Delta\mathbf{m}\right)^2\right), \tag{14}$$

which forces $d=\mu_0-\mu_1$ when the taking limit $\gamma\to\infty$. The combination of (13) and (14) will lead to a posterior distribution over $\mathbf{w}$ that depends on γ and on the observed value d. Taking the limit (Appendix B) gives another Gaussian

$$p(\,\mathbf{w}|\,\mathbf{y},\mathbf{X},d)=N\left(\mathbf{w}\,|\bar{\mathbf{w}}\,,\Sigma_{\bar{\mathbf{w}}}\right)$$

with

$$\bar{\mathbf{w}}=\frac{d\mathbf{B}^{-1}\Delta\mathbf{m}}{\Delta\mathbf{m}^{\mathrm{T}}\mathbf{B}^{-1}\Delta\mathbf{m}}$$

and

$$\Sigma_{\overline{\mathbf{w}}}=\mathbf{B}^{-1}-\frac{\mathbf{B}^{-1}\Delta\mathbf{m}\Delta\mathbf{m}^{\mathrm{T}}\mathbf{B}^{-1}}{\Delta\mathbf{m}^{\mathrm{T}}\mathbf{B}^{-1}\Delta\mathbf{m}}.$$

As would be expected, when an improper prior is used (*i.e.* $\mathbf{A}=\lim_{\alpha\to\infty}\alpha\mathbf{I}$), the mean of this posterior coincides with the maximum likelihood solution given in Section 2.3. Thus our Bayesian discriminant provides a regularised form of the standard Fisher's discriminant where regularisation is provided by the matrix $\mathbf{A}$. Also note that $\Sigma_{\overline{\mathbf{w}}}$ is a positive semidefinite matrix and therefore not invertible. This is a consequence of the fact that any vector $\mathbf{w}$ which does not satisfy the constraint imposed by the distribution $p\left(d|\,\mathbf{w},\mathbf{m}_0,\mathbf{m}_1\right)$ has a posterior probability of zero. Nevertheless, variances associated with output points can still be computed by applying

$$\mathrm{var}\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}\right)=\mathbf{x}^{\mathrm{T}}\mathbf{B}^{-1}\mathbf{x}-\frac{\mathbf{x}^{\mathrm{T}}\mathbf{B}^{-1}\Delta\mathbf{m}\Delta\mathbf{m}^{\mathrm{T}}\mathbf{B}^{-1}\mathbf{x}}{\Delta\mathbf{m}^{\mathrm{T}}\mathbf{B}^{-1}\Delta\mathbf{m}}, \tag{15}$$

which will be zero if the point $\mathbf{x}$ is on the direction of $\Delta\mathbf{m}$.

The Bayesian approach we have outlined leads to a posterior distribution over projections which can be used to compute expected outputs and their associated variances for any given input $\mathbf{x}$. However the limitation imposed by applying a linear model is a strong one. There is an extensive amount of literature explaining why linear models are not always convenient. A common solution is to use a set of nonlinear basis functions such that the new function is linear in the parameters but nonlinear in the input space $f\left(\mathbf{x}\right)=\mathbf{w}^T\phi\left(\mathbf{x}\right)$, [15, 1] . However the problem is shifted to specifying which and what number of basis functions to use. In the next section we shall consider the alternative approach of placing a prior directly over the function $\mathbf{f}$, such that we will be working with a possibly infinite amount of basis functions. This approach will lead to a regularised version of the kernel Fisher discriminant. The probabilistic interpretation of this model will also lead to a principled approach to the selection of kernel parameters.

3.2 Gaussian Process Formulation

Gaussian Processes (GP's) are sub-branch of stochastic processes specified by giving the probability distribution over a finite set of observations $\{f_1, f_2, \ldots, f_N\}$. The specification is given only in terms of their mean vector and covariance matrix. For our convenience, GP's can also be seen as an upper level generalisation of Bayesian regression. In linear regression, a function $f(\mathbf{x})$ is learned by inferring the distribution of a random variable $\mathbf{w}$ and then by applying the deterministic relation $f(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x})$. Instead, in a Gaussian process [13, 19] a prior is placed directly over the function such that a posterior distribution over it can be inferred. Although there are many GP's with an equivalent 'weight space' prior, there exists a large class of them for which no finite dimensional expansion exists. In this regard, a covariance function (or kernel) measures *a priori* the expected correlation between any two pair of points $\mathbf{x}^{(n)}$ and $\mathbf{x}^{(m)}$ in the training set. For example, in a function parameterised as

$$f_n = \mathbf{w}^{\mathrm{T}} \phi\left(\mathbf{x}^{(n)}\right),$$

with a prior over $\mathbf{w}$ specified by a spherical Gaussian with zero mean, $p(\mathbf{w}) = N\left(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}\right)$, the implied correlation between two points is

$$E\left[f_n f_m|\mathbf{w}\right] = \alpha^{-1} \phi\left(\mathbf{x}^{(n)}\right)^{\mathrm{T}} \phi\left(\mathbf{x}^{(m)}\right).$$

In other words, if the feature vector becomes of infinite length the correlation between the two points will lead to a Mercer kernel [16]. However, under these circumstances it no longer makes sense to talk about a prior over the vector $\mathbf{w}$ which would also be infinite. Instead priors over instantiations of the functions are considered.

Therefore, in this subsection we look to reformulate the proposed Bayesian scheme in terms solely of the projected data $\mathbf{f}$. In order to do so, we recall the likelihood we previously proposed (3). Substituting the targets in terms of the class centers, i.e. $\mathbf{t}_q = c_q \mathbf{y}_q$ and furthermore, substituting the values of the centers by its most probable ones, $\hat{c}_q$, will lead to a likelihood that depends only on the projected values $\mathbf{f}$. Some algebraic manipulation (shown in Appendix A) allows us to rewrite the likelihood as

$$p\left(\mathbf{t}|\mathbf{f}, \beta^{-1}\right) \propto \exp\left(-\frac{\beta}{2}\mathbf{f}^{\mathrm{T}}\mathbf{L}\mathbf{f}\right). \tag{16}$$

We follow again the analogy with Bayesian regression presented at the beginning of section (2) but this time from a Gaussian process perspective. In this paradigm, we are interested in making predictions over a new observed value f_* given a set of observations $\mathbf{f}$. In order to do so, we introduce a GP prior formed by the augmented vector $\mathbf{f}_+ = \left[\mathbf{f}^{\mathrm{T}}\ f_*\right]^{\mathrm{T}}$ and an increased kernel $\mathbf{K}_+$, such that

$$p(\mathbf{f}_+) \propto \exp\left(-\frac{1}{2}\mathbf{f}_+^{\mathrm{T}}\mathbf{K}_+^{-1}\mathbf{f}_+\right). \tag{17}$$

The matrix $\mathbf{K}_+$ is of dimensions $(n+1) \times (n+1)$ and has been partitioned as

$$\mathbf{K}_+ = \begin{pmatrix} \mathbf{K} & \mathbf{k} \\ \mathbf{k}^T & k_* \end{pmatrix},$$

with an n column vector $\mathbf{k} = \{k_n | \, i = 1 \dots N\}$ and scalar k_*; the kernel function defined as $k_\times = \mathbf{K}(\mathbf{x}_*, \mathbf{x}_\times)$. Inclusion of the test point will enable the model to make predictions outside the training set. The posterior distribution $p(\mathbf{f}_+ | \, \mathbf{t}, \mathbf{y}, \mathbf{X})$ is readily obtained by combining (16) and (17).

In (3.1) we commented on the risks of learning with the model provided solely with the posterior over $\mathbf{w}$. The Gaussian process framework is no different in this regard, the distance of the projected class means must still lie apart by a fixed distance. We consider the alternative definition of projected class means $\mu_q = \frac{1}{N_q} \mathbf{y}_q^{\mathrm{T}} \mathbf{f}$ given in (5) and rewrite the distribution (14) in terms of $\mathbf{f}$, such that

$$p(d|\mathbf{f}) = \lim_{\gamma \to \infty} \frac{\gamma^{\frac{1}{2}}}{\sqrt{2\pi}} \exp\left(-\frac{\gamma}{2} \left(d - \mathbf{f}^T \Delta\hat{\mathbf{y}} \right)^2 \right), \tag{18}$$

with $\Delta\hat{\mathbf{y}} = N_0^{-1}\mathbf{y}_0 - N_1^{-1}\mathbf{y}_1$. Incorporating the constraint (18) to the posterior will give a distribution of the form $p(\mathbf{f}_+ | \, \mathbf{y}, \mathbf{X}, d)$. We now look to marginalise the vector of observed projections $\mathbf{f}$ in order to obtain the posterior predictive distribution. This marginalisation will once more lead to a distribution dependent on γ, which by taking the limit $\gamma \to \infty$ (Appendix C) will yield a Gaussian of the form

$$p(f_* | \mathbf{y}, \mathbf{X}, d) = N\left(f_* \, | \bar{f}_* \, , \sigma_*^2 \right), \tag{19}$$

with mean and variance

$$\bar{f}_* \propto d\mathbf{k}^{\mathrm{T}} \left(\mathbf{KLK} + \beta^{-1}\mathbf{K} \right)^{-1} \mathbf{K}\Delta\hat{\mathbf{y}}, \tag{20}$$

and

$$\sigma_*^2 = k_* - \mathbf{k}^{\mathrm{T}} \left(\mathbf{K} + \beta^{-1} \left[\mathbf{K}^{-1} + \mathbf{D}^{-1} \right] \right)^{-1} \mathbf{k}, \tag{21}$$

respectively.

The constant of proportionality of (20) and matrix $\mathbf{D}$ (21) are specified in Appendix C .

It is interesting to see that the predictive mean is given by a linear combination of the observed labels, in this case expressed in terms of $\Delta\hat{\mathbf{y}}$. We notice as well that the variance of the prediction is composed by two terms, one representing the observed training data and the other representing a variance purely assigned to the test point. This results are then highly similar to those of typical GP regression, [19, 8]. In the next section we show how this model is equivalent to a regularised version of kernel Fisher 's discriminant .

3.3 Equivalence to KFD

As mentioned before, kernel algorithms rely on expanding the vector $\mathbf{w}$ in the span of the training samples, i.e. $\bar{\mathbf{w}} = \sum_{i=1}^{N} \alpha_i \phi\left(\mathbf{x}^{(i)}\right)$, such that the projection of

a new test point can be written in terms of the inner product

$$\bar{f}_* = \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{k}. \tag{22}$$

In the context of KFD, the vector $\boldsymbol{\alpha}$ is recognised as the eigenvector with maximal eigenvalue that solves the problem posed by formulating Rayleigh's coefficient in the feature space $\mathcal{F}$. In other words, given

$$J(\boldsymbol{\alpha}) = \frac{\boldsymbol{\alpha}^T\mathbf{M}\boldsymbol{\alpha}}{\boldsymbol{\alpha}^T\mathbf{N}\boldsymbol{\alpha}}$$

with $\mathbf{M} = (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^T$, $\mathbf{N} = \mathbf{KLK}$ and $\boldsymbol{\mu}_q = N_q^{-1}\mathbf{K}\mathbf{y}_q$. The solution can be computed by either solving a generalised eigenproblem or by taking

$$\boldsymbol{\alpha} \propto \mathbf{N}^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1). \tag{23}$$

We now resort to a definition of the projected means given in [16], and substitute it in the resulting mean of our model (20). Hence a comparison between (20) and (22) leads to

$$\boldsymbol{\alpha} \propto \left(\mathbf{N} + \beta^{-1}\mathbf{K}\right)^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1). \tag{24}$$

We observe that this is a regularised version of the KFD solution (23). In fact taking the limit $\beta \to \infty$ will give an identical solution, see for example [16]. As the rank of matrix $\mathbf{N}$ is $(N-2)$, making it non invertible, [11] has suggested to solve this deficiency by adding a regulariser (a multiple of the kernel or identity matrix) which in our case is given by $\beta^{-1}\mathbf{K}$.

In our formulation the regularisation term has arised naturally, moreover, our idea of using a likelihood function leads the way to propose a principled approach to estimate the values of the 'regularisation' coefficient as well as the parameters of the kernel.

3.4 Interpreting β

Section 2.2 was devoted to understand the similarities between the likelihood we proposed and the maximisation of Rayleigh's coefficient. More specifically, we were able to determine that the problem simplifies to the point of adjusting the noise level β if the averages of the projected class means lie apart at a fixed distance. This intuition makes sense either for the likelihood proposed (4) and in the constrained version of Rayleigh's coefficient (8):

- Solving the ML problem in (2.2) involved estimating the location of the centers of both distributions and adjusting the precision β. Furthermore, it was demonstrated that the most probable location of each center was μ_q, the projected sample mean. Therefore by fixing $d = \mu_0 - \mu_1$ the number of variables involved in the maximisation is reduced to one, β.

– In a similar way, substituting the constraint making $d = \mu_0 - \mu_1$ into (1) implies that maximisation of Rayleigh's coefficient will only be a function of the variance around each projected class mean. We can see then from equation (6) that such variances are closely related to β.

In order to understand better the physical significance of β, we first rewrite the Rayleigh's coefficient derived under the maximum likelihood framework

$$J = \frac{\hat{\beta} d^2}{N}$$

and we plot the generalisation error as shown in Figure 1.

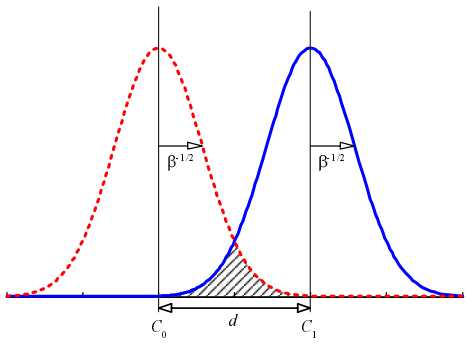


Figure 1. Generalisation error as it relates to β and d. The shaded area gives the generalisation error if the true densities conform to those given by the two Gaussians.

From this figure, we can see that for fixed d, the generalisation error will decrease as β increases, *i.e.* as $\beta^{-1/2}$ decreases. Furthermore, the maximum likelihood solution for the precision obtained in (6) allows us to think that β can be modelled through Bayesian inference. If that is the case, placing a prior distribution over this variance will be equivalent to placing a prior distribution over Rayleigh's coefficient and over the generalisation error. Consider, for example, the case where $d = 2$ and the class priors are equal: if the data does truly map to the mixture distribution with which we are modelling it, then the generalisation error will be

$$E_{\text{eq}} = \frac{1}{2} - \frac{1}{2}\text{erf}\left(\sqrt{\frac{\beta}{2}}\right). \tag{25}$$

If we then consider a gamma distribution as a prior,

$$p(\beta) = \frac{b^a}{\Gamma(a)} \beta^{a-1} \exp(-b\beta),$$

then the MAP solution for β is (Appendix D),

$$\hat{\beta}_{\mathrm{MAP}} = \frac{N + 2a - 2}{\sigma_1^2 + \sigma_0^2 + 2b}. \tag{26}$$

By setting $a = b = 0.5$ we indirectly obtain a uniform distribution over E_{eq}, this important special case leads to a new update equation of the form

$$\hat{\beta}_{\mathrm{MAP}} = \frac{N - 1}{\sigma_1^2 + \sigma_0^2 + 1}, \tag{27}$$

which can be viewed as a regularised version of (6). The prior can also be used to bias β toward low or high generalisation errors if this is thought appropriate.

3.4.1 Other Special Cases Taking the limit as $\beta \to 0$ makes the mean prediction for f_* and its variance take on a much simpler form,

$$\bar{f}_* = \boldsymbol{\alpha}_0^{\mathrm{T}} \mathbf{k}$$

where

$$\boldsymbol{\alpha}_\beta = \frac{d\left(\hat{\mathbf{y}}_0 - \hat{\mathbf{y}}_1\right)}{\left(\hat{\mathbf{y}}_0 - \hat{\mathbf{y}}_1\right)^{\mathrm{T}} \mathbf{K} \left(\hat{\mathbf{y}}_0 - \hat{\mathbf{y}}_1\right)},$$

and

$$\sigma_*^2 = k_*.$$

This result is remarkable for the absence of any requirement to invert the kernel matrix, which greatly reduces the computational requirements of this algorithm. Driving β to zero nullifies the influence of the likelihood, in other words all this model is doing is placing a constraint on the distances between the means given a prior distribution. It has already been pointed out (see *e.g.* [16]) that such a constraint leads to the well known Parzen windows classifiers (sometimes known as probabilistic neural networks) [3].

As we discussed in Section 3.3, taking the limit as $\beta \to \infty$ leads to the standard kernel Fisher's discriminant. From Figure 1 it can be seen that an *a priori* setting of β^{-1} to zero is equivalent to assuming that we can achieve a generalisation error of zero.

4 Optimising Kernel Parameters

One key advantage to interpreting the kernel Fisher 's discriminant as a Gaussian process is that it leads to a principled approach to determining the parameters of the kernel, including the regularisation term. The approach taken is to optimise the normalisation constant in (11), sometimes referred to as the marginal likelihood. We look to optimise

$$L\left(\boldsymbol{\theta}\right) = \log p\left(\mathbf{t}|\mathbf{X}, \mathbf{y}, \beta, \boldsymbol{\theta}\right),$$

with respect to the parameters of our kernel, $\boldsymbol{\theta}$. Recall in Section 2.2 that we optimised the likelihood with respect to the parameters c_0 and c_1 leading to a new encoding of the targets

$$\mathbf{t}_q = \left(\frac{\mathbf{f}^{\mathrm{T}} \mathbf{y}_q}{N_q} \right) \mathbf{y}_q.$$

We back substituted these values in to the likelihood in order to demonstrate the equivalence with maximisation of Rayleigh's coefficient. Unfortunately, one side effect of this process is to cause the target values to become dependent on the inputs and the problem that arises now is that the targets will shift as the kernel parameters are estimated. One solution could be to iterate between determining $\mathbf{t}_0$ and $\mathbf{t}_1$ and optimising the kernel parameters. This approach is simple, but may be difficult to prove convergence properties. We therefore prefer to rely on an expectation-maximisation (EM) algorithm [2] which finesses this issue and for which convergence is proved.

4.1 EM Algorithm

Consider that the log likelihood can be written

$$L(\boldsymbol{\theta}) = \int p(\mathbf{f}|\mathbf{t}, \mathbf{X}, \mathbf{y}, \beta, \boldsymbol{\theta}) \log \frac{p(\mathbf{t}|\mathbf{f}, \mathbf{y}, \beta) \, p(\mathbf{f}|\mathbf{X}, \boldsymbol{\theta})}{p(\mathbf{f}|\mathbf{t}, \mathbf{X}, \mathbf{y}, \beta, \boldsymbol{\theta})} d\mathbf{f}.$$

It is straightforward to show through Jensen's inequality that a lower bound on the likelihood is given by

$$L(\boldsymbol{\theta}) \geq \int q(\mathbf{f}) \log \frac{p(\mathbf{t}|\mathbf{f}, \mathbf{y}, \beta) \, p(\mathbf{f}|\mathbf{X}, \boldsymbol{\theta})}{q(\mathbf{f})} d\mathbf{f}, \tag{28}$$

where $q(\mathbf{f})$ is assumed not to depend on $\boldsymbol{\theta}$. Clearly this lower bound is maximised for $q(\mathbf{f}) = p(\mathbf{f}|\mathbf{t}, \mathbf{X}, \mathbf{y}, \beta, \boldsymbol{\theta})$ when the equality holds. We propose to use an EM algorithm and alternate between optimising the bound with respect to $q(\mathbf{f})$ as E-step and optimising (28) with respect to $\boldsymbol{\theta}$ as M-step. Since the expectation step makes (28) equal to the log-likelihood and the maximisation step optimises this bound, the two steps in tandem are guaranteed to find a (local) maximum for $L(\boldsymbol{\theta})$. The E-step in our model simply involves setting

$$q(\mathbf{f}) \propto \exp\left(-\frac{1}{2} \mathbf{f}^{\mathrm{T}} \Sigma^{-1} \mathbf{f} \right),$$

where

$$\Sigma = \left(\mathbf{K}^{-1} + \beta \mathbf{L} \right)^{-1} \tag{29}$$

and the M-step requires maximisation of

$$\mathcal{L}(\boldsymbol{\theta}) = \langle \log p(\mathbf{f}|\mathbf{X}, \boldsymbol{\theta}) \rangle_{q(\mathbf{f})}, \tag{30}$$

with respect to $\boldsymbol{\theta}$ which can be done with gradient based methods. The notation $\langle \cdot \rangle_{p(x)}$ indicates an expectation under the distribution $p(x)$. Note that for the

updates of the kernel parameters, we have chosen not to apply the constraint on the class means imposed by (18), *i.e.* the kernel we learn is valid for any given d. This is reasonable as the quality of the classification should not be dependent on a given value for d (which can be chosen arbitrarily).

4.1.1 Updating β In the context of the proposed EM updates, it is clear that adjusting the kernel parameters at every iteration will require setting β in some way. To develop an update we again make use of the lower bound and the average distance constraint, as the posterior distribution requires it. We must remember that inclusion of the constraint is necessary as the value $\hat{\beta}$ depends on d. This leads to an update equation

$$\hat{\beta} = \frac{N}{\bar{\sigma}_1^2 + \bar{\sigma}_0^2} \tag{31}$$

where $\bar{\sigma}_1^2 = \sum y_n \left\langle (f_n - \mu_1)^2 \right\rangle$ and the expectation $\langle \cdot \rangle$ is computed under the predictive distribution for the nth training point (19). An expression for $\bar{\sigma}_0^2$ is given in a similar way.

As discussed in Section 3.4 we can also seek a MAP solution and in our experiments we preferred the update

$$\hat{\beta}_{\text{MAP}} = \frac{N-1}{\bar{\sigma}_1^2 + \bar{\sigma}_0^2 + 1}$$

which arises from a uniform prior over the expected generalisation error. Hence the update of $\hat{\beta}$ can be combined with the EM algorithm given above and iterated until convergence of $\hat{\beta}$ and/or $L(\boldsymbol{\theta})$ as outlined in Algorithm 1.

Algorithm 1 A possible ordering of the updates.

Select Convergence tolerances η_β and $\eta_{\boldsymbol{\theta}}$.
Set Initial values $\boldsymbol{\theta}$ and $\hat{\beta}$.
Require data-set $\mathbf{X}$, $\mathbf{y}$.
while change in $\hat{\beta} < \eta_\beta$ and change in $\boldsymbol{\theta} < \eta_{\boldsymbol{\theta}}$ **do**
- Compute kernel matrix $\mathbf{K}$ using $\boldsymbol{\theta}$.
- Update Σ using (29).
- Use scale conjugate gradients to maximise (30) with respect to $\boldsymbol{\theta}$.
- Update $\hat{\beta}$ using (31).

end

5 Experiments

One generally accepted way of determining kernel parameters is to select a range of possible values and cross validate them. This works well when the kernel is

only dependent on one or two parameters but if there are more an alternative approach is required. As it has been suggested, our EM approach gives a good theoretical grounding to perform such estimation. In practice, however, we know that the optimisation routine might get stuck into local minima close to the point of initialisation. To deal with this in a principled way, we first trained our model with different sets of initial parameters and then selected the model with the highest marginal likelihood. This process will be exemplified with experiments done on toy and real data sets.

5.1 Toy data

We constructed three artificial experiments (**ard**, **bumpy** and **overlap**) and used the two-spiral data set from [5] to demonstrate our approach. For each of the experiments we used the update orderings outlined in Algorithm 1. The data sets were designed to test different facets of the following kernel

$$k\left(\mathbf{x}_i, \mathbf{x}_j\right) = \theta_1 \exp\left(-\frac{\theta_2}{2}\left(\mathbf{x}_i - \mathbf{x}_j\right)^{\mathrm{T}} \Theta \left(\mathbf{x}_i - \mathbf{x}_j\right)\right) + \theta_3 \mathbf{x}_i^{\mathrm{T}} \Theta \mathbf{x}_j + \theta_4 + \theta_5 \delta_{ij}, \quad (32)$$

where δ_{ij} is the Kronecker delta and the matrix $\Theta = \mathrm{diag}\left([\theta_6 \ldots \theta_{6+K}]\right)$, with K being the dimension of $\mathbf{x}$. The kernel has three main components: an RBF part ($\theta_1, \theta_2, \Theta$), a linear part ($\theta_3$, Θ), a bias term (θ_3) and white noise (θ_5).

We decided to train each data set three different times by keeping all parameters initialised to 1, except θ_2 which was set to (1, 10, 100) for **ard**, **bumpy** and **overlap** and (1, 500, 5000) for **two-spiral**. We then selected the model that produced the highest marginal likelihood. In all our simulations, we let the algorithm converge whenever the change in a β update was less than 1×10^{-6} or the change in $\boldsymbol{\theta}$ was smaller than 1×10^{-6}. The selected models for each set are summarised in Figure 2.

Experiment A: only one input direction is significant in determining class, but the data is not linearly separable. Experiment B: the spiral data is highly non-linear and requires information from both inputs. Experiment C: the two crossed Gaussian distributions require a non linear decision boundary. Experiment D: the overlapping Gaussians only require a linear decision boundary and information from one input direction. The inferred decision boundaries are also given in Figure 2. The learnt kernel parameters are summarised in Table 1. For each experiment $\hat{\beta}$ was initialised to 1. In all cases we obtained arguably good solutions. Irrelevant directions were down-weighted and the relative use of non-linear *vs* linear decision boundaries (as indicated by the ratio of θ_2 to θ_3) was as expected.

Figure 3 below shows an example of the result of training **two-spiral** with a bad initialisation. Notice the value of the marginal likelihood in this case is smaller to the one presented in Figure 2. The kernel parameters determined by the algorithm for the four experiments are given in Table 1.

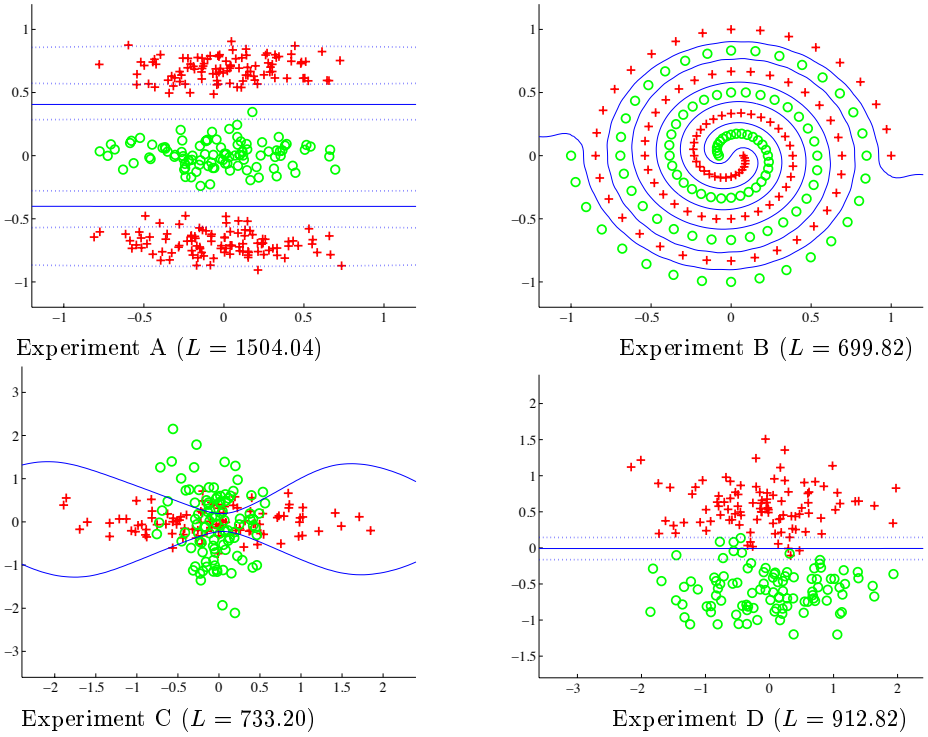


Figure 2. Results from the four experiments. Two classes are shown as pluses and circles. The decision boundary is given by the solid line. Dotted lines indicate points at 1/4 of the distance (as measured in the projected space) from the decision boundary to the class mean. Log-likelihood value appears enclosed by brackets. .

6 Benchmark data sets

In order to evaluate the performance of our approach, we performed a series of experiments with well known data sets. We used the synthetic set **banana** provided by Gunnar Rätsch [2] and 10 other real world data sets from the **UCI**, **DELVE** and **STATLOG** benchmark repositories[3]. Whenever applicable, the experimental setup was chosen according to [14] and [11], where (among others) the compared methods were: **RBF**, a single RBF classifier; **AB(r)**, regularised AdaBoost; **SVM**, a Support Vector Machine (Gaussian kernel) and **KFD**, the

[2] Data sets can be obtained from at http://mlg.anu.edu.au/~raetsch. Due to computational reasons, in this work we did not implement our approach on the data sets Flare-Solar, Image and Splice because they exceeded 1000 training points.

[3] The **breast cancer** domain was obtained from the University Medical Center, Institute of Oncology, Ljubljana, Yugoslavia. Thanks to M. Zwitter and M. Soklic for the data.

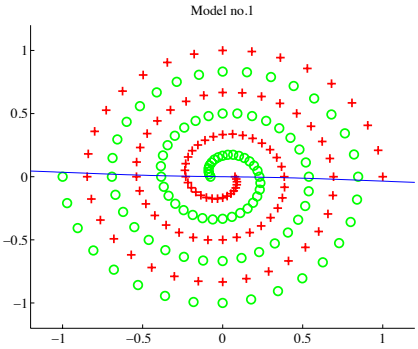


Figure 3. The solution for the spiral data with a poor initialisation $\theta_2 = 1$. Associated log-likelihood $L = 605.93$.

Experiment	$\theta_1 \times 10^{-5}$	θ_2	θ_3	θ_4	$\theta_5 \times 10^{-3}$	$(\theta_6 \times 10^{-4}$,	$\theta_7 \times 10^{-2})$
A	4.74	589.11	1.89×10^{-5}	2.74	4.40	1.62	3.76
B	9.32	9230.00	1.88×10^{-4}	2.75	8.10	323.00	2.92
C	2.05	29.68	4.32×10^{-7}	2.75	5.10	0.02	11.91
D	0.19	13.39	1.10×10^{-3}	2.75	0.99	2.88	2.90

Table 1. The parameters learnt for the different experiments. Some columns are scaled to ease comparisons, scaling factors are given in the top row.

kernel Fisher discriminant (in [11]). Data was processed to have binary classes and then partitioned into 100 test and training sets, see [14] for further details. As mentioned earlier, non Bayesian schemes resort to cross-validation to estimate model parameters. In all these methods, the parameters required were estimated by running 5-fold cross validation on the first 5 realizations of the training sets. The selected parameters were then chosen as the median over these 5 estimates.

In our first approach, denoted by **BFD**, we used 5 different initialisations of the 'inverse width' part of an RBF kernel to train our models. Following the previous scheme, we trained on the first 5 realisations of each data set and computed their corresponding marginal loglikelihood. In this way, a matrix of 5×5 elements was obtained with each element containing a marginal likelihood. Furthermore, for each realisation we selected the model with highest associated $L(\boldsymbol{\theta})$ (a vector of length 5 was formed). The parameters were selected from the median of the inverse widths associated to each likelihood in the vector that was described. Under this framework the regularisation β was considered a parameter as well.

The experiments with **BFD** show, in general, comparable results with the previous approaches. However, we can take a step further and take advantage of some of the features provided by the Bayesian framework of inference. More specifically, we tested the Automatic Relevance Determination feature that

might be obtained from selecting an appropriate prior [9]. In order to do it, we performed the same classification experiments but with a kernel of the form described in (32). A kernel of this nature would have been very difficult to implement with the cross validated methods as there are too many parameters to explore. We present the results of these new experiments under the column **BFD**ARD in Table 2. All experiments were trained on the 5 realisations of data and with θ_2 initialised to one. The algorithm was stopped when the change in $\beta < 1 \times 10^{-4}$ and/or a change in $L(\boldsymbol{\theta}) < 1 \times 10^{-4}$. We observe a substantial improvement for some of the data sets, compared with **BFD**, implying that ARD prior can help.

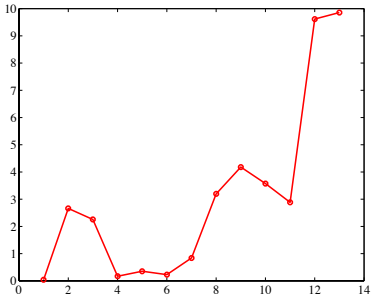
The selection of an 'ARD' kernel (32) implies that irrelevant features are down weighted and effectively pruned out of the model. This effect could be visualised on some of the examples on the synthetic data sets and we would expect to see it to some extent on the experiments with real data. Table (3) shows the plots of the values of the weight vector Θ for Heart and Thyroid. In the first case, the attributes 1, 4, 5 and 6 were pruned out of the discrimination process; whereas in the second case the analogous effect takes place for the features: 1, 3 and 5 were removed.

Dataset	**RBF**	**AB(r)**	**SVM**	**KFD**	**BFD**	**BFD**ARD
Banana	10.8 ± 0.6	*10.9 ± 0.4*	11.5 ± 0.7	$\mathbf{10.8 \pm 0.5}$	11.5 ± 0.7	12.4 ± 0.9
B. Cancer	27.6 ± 4.7	26.5 ± 4.5	26.0 ± 4.7	*25.8 ± 4.6*	28.8 ± 4.4	$\mathbf{24.1 \pm 4.6}$
Diabetes	24.3 ± 1.9	23.8 ± 1.8	*23.5 ± 1.7*	$\mathbf{23.2 \pm 1.6}$	27.2 ± 2.4	24.7 ± 1.8
German	24.7 ± 2.4	24.3 ± 2.1	*23.6 ± 2.1*	23.7 ± 2.2	$\mathbf{23.4 \pm 0.2}$	25.6 ± 2.3
Heart	17.6 ± 3.3	16.5 ± 3.5	*16.0 ± 3.3*	16.1 ± 3.4	16.1 ± 3.3	$\mathbf{15.9 \pm 3.5}$
Ringnorm	1.7 ± 0.2	*1.6 ± 0.1*	1.7 ± 0.1	$\mathbf{1.5 \pm 0.1}$	1.8 ± 0.4	1.7 ± 0.2
Thyroid	*4.5 ± 2.1*	4.6 ± 2.2	4.8 ± 2.2	$\mathbf{4.2 \pm 2.1}$	5.4 ± 2.4	4.6 ± 2.3
Titanic	23.3 ± 1.3	*22.6 ± 1.2*	$\mathbf{22.4 \pm 1.0}$	23.2 ± 2.0	24.7 ± 2.0	$\mathbf{22.4 \pm 0.3}$
Twonorm	2.9 ± 0.3	2.7 ± 0.2	3.0 ± 0.2	*2.6 ± 0.2*	$\mathbf{2.4 \pm 0.1}$	*2.6 ± 0.1*
Waveform	10.7 ± 1.1	$\mathbf{9.8 \pm 0.8}$	*9.9 ± 0.4*	*9.9 ± 0.4*	15.0 ± 0.8	14.2 ± 0.3

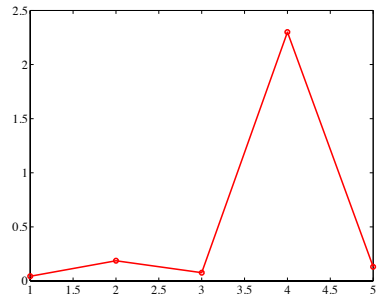
Table 2. Results of experiments with real data sets. **BFD** refers to our model applied with a standard one dimensional RBF kernel, whereas **BFD**ARD indicates an ARD-RBF kernel.

7 Conclusions and future work

We have presented a Bayesian approach to discriminant analysis that corresponds to kernel Fisher's discriminant. Regularisation of the discriminant arises naturally in the Bayesian approach and through maximising the marginal likelihood we are able to determine kernel parameters. This paper has established the theoretical foundations of the approach and has shown that for a range of simple toy problems the methodology does discover sensible kernel parameters.



Heart



Thyroid

Table 3. Plots of the weights that were learnt for Heart and Diabetes data sets. We used the kernel specified in (32).

The optimisation is only guaranteed to find local minimum and therefore the quality of the solution can be sensitive to the initialisation. We performed experiments on real world data obtaining results which are competitive with the state of the art, moreover, we were able to do some relevance determination on the data set features.

Future directions of this work will be centered on sparsifying the kernel matrix. We intend to adapt the Informative Vector Machine model to our framework [7]. This should make larger data sets practical. At present, we are restricted by the $\mathcal{O}\left(N^3\right)$ complexity associated with inverting the kernel matrix. Another direction of research will consist of allowing the model to learn in the presence of label noise building on work in [6].

Acknowledgements

Both authors gratefully acknowledge support from the EPSRC grant number GR/R84801/01 'Learning Classifiers from Sloppily Labelled Data'.

Bibliography

[1] C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1st edition, 1995.

[2] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39:1–38, 1977.

[3] R. O. Duda and P. E. Hart. *Pattern recognition and scene analysis*. John Wiley, 1st edition, 1973.

[4] I. Joliffe. *Principal Component Analysis*. Springer-Verlag, 1986.

[5] K. Lang and M. Witbrok. Learning to tell two spirals apart. Morgan Kauffman, 1988.

[6] N. D. Lawrence and B. Schölkopf. Estimating a kernel Fisher discrminant in the presence of label noise. In C. E. Brodley and A. P. Danyluck, editors, *Proceedings of the 18th International Conference on Machine Learning*, San Francisco, CA., jul 2001. Morgan Kauffman.

[7] N. D. Lawrence, M. Seeger, and R. Herbrich. Sparse Bayesian learning: the informative vector machine. Technical report, Department of Computer Science University of Sheffield, 2002.

[8] D. J. Mackay. Introduction to Gaussian processes. Technical report, Department of Physics, University of Cambridge.

[9] D. J. Mackay. Probable networks and plausible predictions - a review of practical Bayesian methods for supervised neural networks. *Network: Computation in Neural Systems*, 6(3):469–505, 1995.

[10] S. Mika. *Kernel Fisher Discriminants*. PhD thesis, Techniseche Universitaat, Berlin, Germany, 2002.

[11] S. Mika, G. Rätsch, B. Schölkopf, and K.-R. Moeller. Fisher discriminant analysis with kernels. In Y.-H. Hu, E. W. J. Larsen, and S. Douglas, editors, *Neural Networks for Signal Processing IX*, pages 41–48. IEEE, 1999.

[12] S. Mika, A. J. Smola, and B. Schölkopf. An improved training algorithm for kernel Fisher discriminants. In T. Jaakkola and T. Richardson, editors, *Proceedings of the Eigth International Workshop on Artificial Intelligence and Statistics*, San Francisco, CA., 2001. Morgan Kauffman.

[13] A. O'Hagan. Curve fitting and optimal design for prediction. *Journal of the Royal Statistical Society*, B(40):1–42, 1978.

[14] G. Rätsch, T. Onoda, and K.-R. Mueller. Soft margins for AdaBoost. Technical Report NC-TR-98-021, Neural Networks and Computational Learning Theory, 1998.

[15] D. Ruppert, M. Wand, and R. Carroll. *Semiparametric regression*. Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, U.K., 2003.

[16] B. Schölkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond*. MIT Press, 2002.

[17] J. A. Suykens and J. Vandewalle. Least squares support vector machines. *Neural Processing Letters*, 9(3), 1999.

[18] V. Vapnik. *The nature of statistical learning.* Springer-Verlag, New York, 1995.

[19] C. K. Williams. *Prediction with Gaussian Processes: from linear regression to linear prediction and beyond*, behavioural and social sciences 11. D. Kluwer, Dordrecht, The Netherlands, 1998.

A Expressing the likelihood in terms of f

Substituting the values $\mathbf{t}_q = c_q \mathbf{y}_q$ into (3) and with some straightforward algebra we can obtain equation (4),

$$p\left(\mathbf{t}\left|\mathbf{f}, \beta^{-1}\right.\right) = \frac{\beta^{N/2}}{(2\pi)^{N/2}} \exp\left\{-\frac{\beta}{2}\left[\sum_{n=1}^{N} y_n \left(c_1 - f_n\right)^2 + \sum_{n=1}^{N}\left(1-y_n\right)\left(c_0 - f_n\right)^2\right]\right\}. \tag{33}$$

A further substitution of the most probable values for the locations $\hat{c}_q$ and considering that $y_n = y_n^2$ will lead into the desired expression. Working with the argument of the exponential function in (33) will lead to

$$\begin{aligned}
&\sum_{n=1}^{N} y_n^2 \left(\frac{1}{N_1}\mathbf{y}_1^T\mathbf{f} - f_n\right)^2 + \sum_{n=1}^{N}\left(1-y_n\right)^2 \left(\frac{1}{N_0}\mathbf{y}_0^T\mathbf{f} - f_n\right)^2 \\
&= \mathbf{f}\mathbf{f}^T - \tfrac{1}{N_1}\mathbf{y}_1^T\mathbf{f}\mathbf{f}^T\mathbf{y}_1 - \tfrac{1}{N_0}\mathbf{y}_0^T\mathbf{f}\mathbf{f}^T\mathbf{y}_0 \\
&= \mathbf{f}^T\mathbf{L}\mathbf{f}.
\end{aligned}$$

B Weight space approach

In order to derive the distribution of $\mathbf{w}$ under the constraint d, we first realise that the combination of $p\left(\mathbf{w}\left|\mathbf{0}, \mathbf{B}^{-1}\right.\right)$ and $p\left(d\left|\mathbf{w}, \mathbf{m}_0, \mathbf{m}_1\right.\right)$ yields

$$p(\mathbf{w}|\mathbf{y}, \mathbf{X}, d, \gamma) = \lim_{\gamma\to\infty} N\left(\gamma d \Sigma_\gamma \Delta\mathbf{m}, \Sigma_\gamma^{-1}\right), \tag{34}$$

with $\Sigma_\gamma = \mathbf{B} + \gamma\Delta\mathbf{m}\Delta\mathbf{m}^{\mathrm{T}}$. Inversion of Σ_γ through Morrison-Woodbury formula will allow to take the limit, such as is shown below

$$\Sigma_\gamma^{-1} = \mathbf{B}^{-1} - \frac{\mathbf{B}^{-1}\Delta\mathbf{m}\Delta\mathbf{m}^T\mathbf{B}^{-1}}{\gamma^{-1} + \Delta\mathbf{m}^T\mathbf{B}^{-1}\Delta\mathbf{m}},$$

hence

$$\Sigma_{\bar{\mathbf{w}}} = \lim_{\gamma\to\infty} \Sigma_\gamma^{-1}.$$

Substitution of Σ_γ^{-1} into the mean of (34) along with some straightforward algebra will lead into $\bar{\mathbf{w}}$.

C Gaussian process approach

Distribution of a new point f_*

The distribution $p(f_*, d|\mathbf{y}, \mathbf{X})$ is obtained by computing $\int p\left(\mathbf{f}_+\left|\mathbf{y}, \mathbf{X}_N, d\right.\right) \partial\mathbf{f}$. In order to do so, we resort to partition the inverse of the augmented kernel matrix

$$\mathbf{K}_+^{-1} = \begin{pmatrix} \mathbf{C} & \mathbf{c} \\ \mathbf{c} & c_* \end{pmatrix},$$

with

$$\begin{aligned} c_* = &\quad \left[k_* - \mathbf{k}^T\mathbf{K}^{-1}\mathbf{k}\right]^{-1} \\ \mathbf{c} = &\quad -c_*\mathbf{K}^{-1}\mathbf{k} \\ \mathbf{C} = &\quad \mathbf{K}^{-1} + c_*\mathbf{K}^{-1}\mathbf{k}\mathbf{k}^T\mathbf{K}^{-1}. \end{aligned} \tag{35}$$

The result of this integral will lead into

$$p(f_*, d|\mathbf{y}, \mathbf{X}) \propto \exp\left\{-\frac{1}{2\sigma_*^2}\left(f_* - \bar{f}_*\right)^2 + \frac{1}{2}\left[(\gamma d)^2\, \Delta\hat{\mathbf{y}}\mathbf{P}\Delta\hat{\mathbf{y}}^{\mathrm{T}} - \gamma d^2\right]\right\},$$

with

$$\mathbf{P} = \left(\mathbf{C} + \beta\mathbf{L} + \gamma\Delta\hat{\mathbf{y}}\Delta\hat{\mathbf{y}}^{\mathrm{T}}\right)^{-1}$$

and

$$\bar{f}_* = \lim_{\gamma\to\infty} -\gamma d\left(c_* - \mathbf{c}^T\mathbf{P}\mathbf{c}\right)^{-1}\mathbf{c}^T\mathbf{P}\Delta\hat{\mathbf{y}} \tag{36}$$

$$\sigma_*^2 = \lim_{\gamma\to\infty}\left(c_* - \mathbf{c}^T\mathbf{P}\mathbf{c}\right)^{-1}, \tag{37}$$

Derivations of σ_*^2 and f_* follow.

Working out the covariance Substituting (35) into σ_*^2 gives

$$\sigma_*^2 = \lim_{\gamma\to\infty}\left(c_* - c_*^2\mathbf{D}_\gamma\mathbf{k}\right)^{-1}, \tag{38}$$

with

$$\mathbf{D}_\gamma = \mathbf{K}\left(\gamma\Delta\hat{y}\Delta\hat{\mathbf{y}}^{\mathrm{T}} + \beta\mathbf{L}\right)\mathbf{K} + \mathbf{K}. \tag{39}$$

Applying Morrison-Woodbury formula to (38)

$$\begin{aligned} \sigma_*^2 &= \lim_{\gamma\to\infty}\left[c_* - c_*^2\mathbf{k}^T\left(\mathbf{D}_\gamma + c_*\mathbf{k}\mathbf{k}^T\right)^{-1}\mathbf{k}\right]^{-1} \\ &= \lim_{\gamma\to\infty} k_* - \mathbf{k}^T\left[\mathbf{K}^{-1} - \mathbf{D}_\gamma^{-1}\right]\mathbf{k}. \end{aligned}$$

Inversion of $\mathbf{D}_\gamma$ allows to obtain $\mathbf{D}$ by taking the limit,

$$\mathbf{D}_\gamma^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{K}\Delta\hat{\mathbf{y}}\left(\frac{1}{\gamma}\mathbf{I} + \mathbf{K}\Delta\hat{\mathbf{y}}\mathbf{A}\Delta\hat{\mathbf{y}}^T\mathbf{K}\right)^{-1}\Delta\hat{\mathbf{y}}^T\mathbf{K}\mathbf{A}, \text{ s.t.}$$

$$\mathbf{D}^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{K}\Delta\hat{\mathbf{y}}\left(\mathbf{K}\Delta\hat{\mathbf{y}}\mathbf{A}\Delta\hat{\mathbf{y}}^T\mathbf{K}\right)^{-1}\Delta\hat{\mathbf{y}}^T\mathbf{K}\mathbf{A},$$

where $\mathbf{A} = \beta\mathbf{K}\mathbf{L}\mathbf{K} + \mathbf{K}$.

Working out the mean Inverting the scalar inside (36) and substituting the value of c_* gives

$$\begin{aligned}\bar{f}_* = & \lim_{\gamma\to\infty} \gamma d\sigma_*^2 c_* \mathbf{k}^T \left(\mathbf{D}_\gamma + c_* \mathbf{k}\mathbf{k}^T\right)^{-1} \mathbf{K}\Delta\hat{\mathbf{y}} \\ = & \lim_{\gamma\to\infty} \gamma d\sigma_*^2 \left[k_* - \mathbf{k}^T \left(\mathbf{K}^{-1} - \mathbf{D}_\gamma^{-1}\right)\mathbf{k}\right]^{-1} \mathbf{k}^T \mathbf{D}_\gamma^{-1} \mathbf{K}\Delta\hat{\mathbf{y}}\end{aligned}$$

Using the expressions for σ_*^2 along with that of $\mathbf{A}$ will lead into

$$\begin{aligned}\bar{f}_* = & \lim_{\gamma\to\infty} \gamma d\mathbf{k}^T \mathbf{D}_\gamma^{-1} \Delta\bar{\mathbf{k}} \\ = & \lim_{\gamma\to\infty} d\left(\frac{1}{\gamma} + \Delta\hat{\mathbf{y}}^T \mathbf{K}\mathbf{A}^{-1}\mathbf{K}\Delta\hat{\mathbf{y}}\right)^{-1} \mathbf{k}^T \mathbf{A}^{-1}\mathbf{K}\Delta\hat{\mathbf{y}}\end{aligned}$$

and the desired result

$$\bar{f}_* = \frac{d\mathbf{k}^T \mathbf{A}^{-1}\mathbf{K}\Delta\hat{\mathbf{y}}}{\Delta\hat{\mathbf{y}}^{\mathrm{T}}\mathbf{K}\mathbf{A}^{-1}\mathbf{K}\Delta\hat{\mathbf{y}}}.$$

D Obtaining MAP solution over β

Making

$$\begin{aligned}V = & \textstyle\sum_{n=1}^N y_n \left(c_1 - f_n\right)^2 + \sum_{n=1}^N \left(1 - y_n\right)\left(c_0 - f_n\right)^2 \\ = & \quad \sigma_1^2 + \sigma_0^2.\end{aligned}$$

such that the likelihood (4) becomes

$$p\left(\mathbf{t}\,\middle|\,\mathbf{f}, \beta^{-1}\right) = \frac{\beta^{N/2}}{(2\pi)^{N/2}} \exp\left\{-\frac{\beta}{2}V\right\}.$$

Then combining it with a gamma prior $G\left(\beta|\, a, b\right)$ gives a Gamma posterior of the form $G\left(\beta|\, N/2 + a, (V/2 + b)\right)$, that is

$$p\left(\beta|\, \mathbf{t}, \mathbf{f}\right) \propto \beta^{N/2+a-1} \exp\left\{-\beta\left(\frac{V}{2} + b\right)\right\}.$$

Taking the derivative of the log of this distribution and equating it to zero will give (26).