Model selection in high dimensions: a quadratic-risk-based approach

Surajit Ray

Boston University, USA

and Bruce G. Lindsay

Pennsylvania State University, University Park, USA

[Received June 2006. Revised August 2007]

Summary. We propose a general class of risk measures which can be used for data-based evaluation of parametric models. The loss function is defined as the generalized quadratic distance between the true density and the model proposed. These distances are characterized by a simple quadratic form structure that is adaptable through the choice of a non-negative definite kernel and a bandwidth parameter. Using asymptotic results for the quadratic distances we build a quick-to-compute approximation for the risk function. Its derivation is analogous to the Akaike information criterion but, unlike the Akaike information criterion, the quadratic risk is a global comparison tool. The method does not require resampling, which is a great advantage when point estimators are expensive to compute. The method is illustrated by using the problem of selecting the number of components in a mixture model, where it is shown that, by using an appropriate kernel, the method is computationally straightforward in arbitrarily high data dimensions. In this same context it is shown that the method has some clear advantages over the Akaike information criterion and Bayesian information criterion.

Keywords: Global comparison of models; High dimensional data; Mixture models; Model selection; Quadratic distance; Quadratic risk; Spectral degrees of freedom

1. Introduction

In this paper we consider data-based evaluation of statistical models, where by a statistical model we mean a parametric family of distributions. We shall denote a parametric model by \( \mathcal{M} \), and an element of the model by \( M_\theta \), for \( \theta \) in some parameter space \( \Theta \), so that \( \mathcal{M} = \{ M_\theta; \theta \in \Theta \} \). A collection of a set of probability models \( \mathcal{M}_k, k = 1, 2, \ldots \), will constitute a model class, say \( \mathcal{M} \).

We shall focus on the problem of selecting one statistical model from the competing models in the class \( \mathcal{M} = \{ \mathcal{M}_k \} \), where \( k \) is an index for the parametric model that is under consideration. In the example that we consider in Section 4, \( \mathcal{M} \) is the class of all finite mixtures of multivariate Gaussians and, for each \( k \), \( \mathcal{M}_k \) is the set of all \( k \)-component multivariate Gaussian mixtures. A particular model element \( M_\theta \in \mathcal{M}_k \) is a \( k \)-component mixture with specific values for the parameters \( \theta \). Let us also use \( F_T \) to denote the true distribution of the data.

In general, model selection can be approached from two main philosophical perspectives:

(a) testing-based model assessment and
(b) parsimony-based model assessment.
The testing-based approach involves testing, for each \( k \), \( H_0 : F_\tau \in \mathcal{M}_k \) versus \( H_1 : F_\tau \in \mathcal{M}_{k+1} \). The likelihood ratio test is often used to build a testing-based model selection method. However, we do not favour a simple testing approach, because, in the spirit of George Box, our prior belief is that every restricted model is flawed, in which case every model would be rejected by testing at some sample size. Under this belief, a suitable model selection theory should instead be based on the adequacy of the model approximation.

In parsimony-based model selection, we define a selection criterion, which usually involves a term accounting for the goodness of fit and a term penalizing richer models (models with more parameters). Then we choose as the best model among a given subset of models, i.e. \( \mathcal{M}_{\text{best}} \in \mathcal{M} \), the model which optimizes our chosen criterion. The information-criterion-based approaches such as the Akaike information criterion (AIC) (Akaike, 1973), the Bayesian information criterion (BIC) (Schwarz, 1978) and other model selection criteria that are based on approximations of Bayes factors (Haughton, 1988) all fall under this large class of parsimony-based model selection tools. These tools are usually designed under a specific probability structure or are designed to achieve specific goals. For example the AIC is designed to achieve maximum predictive accuracy whereas the BIC evaluates the posterior probability of the competing models with specified priors (Yang, 2005).

In practice the most parsimonious model is determined by the model that is associated with minimum risk under the specific loss function. For example, in the case of the AIC, the loss function of the risk is the Kullback–Leibler (KL) distance. We shall here consider model selection from the same basic framework but we shall replace KL distance with quadratic distance (Lindsay et al., 2007). We do so because the family of quadratic distances allows for a very straightforward estimation of risk when the distance itself is used as the loss function. As evident later in this paper, the resulting quadratic risk has many desirable properties especially for high dimensional data and for model comparisons in a non-regular parametric set-up.

Before describing the quadratic risk we want to point out two underlying themes of this paper. First, as our risk measure is in the spirit of the AIC, we shall draw a parallel to the AIC and, wherever necessary, contrast our quadratic risk measures with the AIC. Also, this research was originally motivated by the problem of model selection for high dimensional data in the normal mixture model, a non-regular parametric class. We shall use this important model to demonstrate and address pivotal mathematical and computational problems in model selection.

1.1. Quadratic-risk-based model selection

We shall develop our model selection criteria within the following risk framework. The loss that is incurred in using a model element \( M_\theta \) (from a model \( \mathcal{M} \)) when the true distribution is \( F_\tau \) will be denoted by \( a(F_\tau, M_\theta) \). The function \( a \) should measure in some meaningful fashion the price of using \( M_\theta \) to approximate \( F_\tau \). The risk of using the model \( \mathcal{M} \), given a particular parameter estimator \( \hat{\theta} \), will be defined to be

\[
\rho_a(F_\tau, \mathcal{M}, m) = E\{a(F_\tau, M_\hat{\theta})\},
\]

where the \( m \) on the left-hand side indicates that the estimate \( \hat{\theta} \) on the right-hand side was based on a sample of size \( m \), and the expectation is taken over the true distribution. We alert the reader that we are retaining \( m \) as a risk argument because we shall consider the estimation of risk for values of \( m \) other than the actual data size \( n \). Our goal here is to choose the model that minimizes this risk, noting that the lowest risk model is likely to depend on \( m \), i.e. in this formal definition the problem is not to pick a true model, or even a closest-to-true model, from \( M \): it is to pick a model for which our estimated model element \( M_\hat{\theta} \) will be the most similar to \( F_\tau \), on average, in samples of size \( m \). Although models with many parameters are likely to be closer to the true
model, the estimation of the model parameters creates extra variability that is penalized in the risk calculation.

The AIC model selection criterion arises in this context as follows. If we assume both \( M_\theta \) and \( F_\tau \) to be continuous density functions, with densities \( m_\theta(x) \) and \( f_\tau(x) \) respectively, the KL loss function is

\[
\text{KL}(F_\tau, M_\theta) = \int \log \left( \frac{f_\tau(x)}{m_\theta(x)} \right) f_\tau(x) \, dx.
\]

Using twice the KL function as a loss function, we can split the risk into two pieces:

\[
\rho_{\text{KL}}(M, m) = -2 \mathbb{E} \left[ \int \log \{ m_\hat{\theta}(x) \} f_\tau(x) \, dx + \int \log \{ f_\tau(x) \} f_\tau(x) \, dx \right]
\]

\[
= -2 \mathbb{E} \left[ \int \log \{ m_\hat{\theta}(x) \} f_\tau(x) \, dx \right] + 2 \int \log \{ f_\tau(x) \} f_\tau(x) \, dx.
\]

Here again \( m \) represents the sample size that is used for the estimator \( \hat{\theta} \) and we shall think of it as a variable that changes the magnitude of the penalty function. Of course, the convention is to replace \( m \) with \( n \), the actual sample size. The AIC is based on estimation of the first term on the right-hand side of equation (1). This can be done because the second term does not depend on any of the models under consideration. As a consequence, one does not estimate the full risk of using a model, but only its relative risk, which in turn depends on the other models under consideration, i.e. a model that is chosen as the best among a class of models will still have an uncertain amount of total risk, and so could in fact be a poor fit.

Our quadratic risk is defined by using a quadratic distance (Lindsay et al., 2007) as the loss function \( \cdot, \cdot \). A quadratic distance between the true distribution \( F_\tau \) and any proposed distribution \( G \), where \( G \) can be either discrete or continuous, has the form

\[
d_K(F_\tau, G) = \int \int K_G(s, t) \, d(F_\tau - G)(s) \, d(F_\tau - G)(t),
\]

where \( K_G \) is a kernel of appropriate dimensions (details are given in Section 2), that can be chosen by the user to meet specific goals. Note that we shall allow the kernel to depend on \( G \), the second argument in distance \( d_K(F_\tau, G) \).

A familiar example of this type is the Pearson \( X^2 \)-distance on a partitioned sample space.

### 1.1.1. Example 1.

Let \( A_1, \ldots, A_C \) be a partition of the sample space \( S \) into \( C \) bins and let \( G \) be a target probability measure. Define the kernel by

\[
K_G(x, y) = \sum_{i=1}^{C} \frac{\mathbb{I}[x \in A_i] \mathbb{I}[y \in A_i]}{G(A_i)},
\]

where \( \mathbb{I} \) is the indicator function and \( G(A_i) = \int_{A_i} dG(x) \). The resulting quadratic distance is the Pearson \( X^2 \)-distance, given by

\[
\frac{\sum_{i=1}^{C} \left( F_\tau(A_i) - G(A_i) \right)^2}{G(A_i)}.
\]

We shall be using the Pearson \( X^2 \)-example throughout the paper as a way to exemplify our distance calculations in the simplest possible setting. In the example that we introduce in Section 4 we shall be using a Gaussian kernel.
1.2. Why quadratic-risk-based model selection?

Why do we propose this new measures of model selection? Here we discuss three key issues that played a role in the development of our methodology.

1.2.1. Local versus global comparison of models

Quadratic distance does not separate into two parts in the manner that KL distance does (see equation (1)). This means that we must estimate the absolute risk for a model. In the process, however, we can learn whether the models provide a low level of absolute risk, whereas for the AIC we have no assurance that any model, even the best one, fits well in an absolute sense.

The absolute quadratic risk of a parametric model can be calibrated by comparing its risk with the risk of the empirical distribution function. The latter estimator has no lack of fit but, owing to its flexibility, will have the highest variability. As we shall see in our simulations, its risk provides an excellent benchmark for model quality. In effect, failure to meet this standard means that either more model building is needed or non-parametric approaches should be used. In our application section we provide examples where the global comparison provides crucial decision-making information in model selection.

1.2.2. The role of model regularity conditions

The approximations that are used to arrive at the AIC, BIC and most other information criteria depend strongly on regularity conditions (boundary conditions and nested parameter structure among others) some of which are violated in important model selection exercises. For example, the regularity problems for nested mixture models are well known (see Section 4). Similar problems occur in the structural equations model (Bollen et al., 2006) and the multilevel model, where the boundary between nested models is irregular. All these make the usual asymptotic expansions inappropriate.

In contrast, our method is based on global tests of goodness of fit. As a result the test statistics that are used have asymptotic expansions that do not depend on regularity assumptions for nested models and so have wider validity.

1.2.3. Addressing computational challenges in high dimensions

A primary goal of our research was to devise model selection methods that would be practical for high dimensional problems. However, quadratic distances, such as the one that we are proposing as the loss function of our risk, could require numerical integration of the same dimension as the data vector, which would defeat our purpose. For our mixture example we shall show how to avoid this computational burden by using a ‘rational’ kernel for the distance; this gives the needed integration in a closed form. A detailed description of construction of these kernels can be found in Lindsay et al. (2007).

An additional challenge in high dimensions is the construction of distances whose operating characteristics can be tuned to the dimension of the problem and the sample size. In a $\chi^2$-test this would be done by choosing the number and location of the bins. Lindsay et al. (2007) defined the degrees of freedom of a quadratic distance and discussed its use to select distances. The kernel that we shall use in the mixture example includes a ‘tuning parameter’, that allows us to select suitable degrees of freedom for the distance and also allows us to analyse the data at multiple resolutions.
1.3. Description of paper
Section 2 provides a detailed description of our choice of the quadratic distance in defining the quadratic risk, the appropriateness of such a definition in the context of high dimensional data sets and the estimation of quadratic risk. In Section 3 we build model selection tools based on our quadratic risk. Next, in Section 4, we demonstrate how this model selection tool can be applied to select the number of components in a multivariate normal mixture. Application to real data and simulated data sets and a comparison of our method with existing model selection tools are described in Section 5. Section 6 contains a discussion and Appendix A provides proofs of results that are stated in the paper.

The programs that were used to analyse the data can be obtained from http://www.blackwellpublishing.com/rss

2. Quadratic risk and estimation of quadratic risk
This section provides the foundational details and the estimation techniques for quadratic risk. We shall start by defining the quadratic distance between two arbitrary probability measures. Then, after deriving an unbiased estimator of the quadratic distance we shall use it to build an unbiased cross-validation-based estimator for the quadratic risk measure. Although analytically attractive, this cross-validation method is computationally very expensive whenever parameter estimation is expensive, especially so for large data sets or when the dimensions are high. So, in the latter part of this section we shall derive an AIC-like approximation to the risk measure, which is largely based on the decomposition of our quadratic risk and the asymptotics of quadratic distance. The definitions and results on quadratic distance that are discussed in this section are described in greater detail in Lindsay et al. (2007) and Ray (2003).

2.1. Quadratic distance: definition and empirical estimate
2.1.1. Quadratic distance framework
First we provide the details of construction of the class of quadratic distance, which we use as the loss function of our risk measures. We define $d_K(F_\tau, G)$ to be the distance between the true density $F_\tau$ and any proposed distribution $G$, provided that they are defined on the same sample space $S$. The building block for our distance will be $K(s, t)$, a bounded symmetric kernel function on $S \times S$, which is conditionally non-negative definite ($K$ is conditionally non-negative definite if $\int \int K(s, t) \, d\sigma(s) \, d\sigma(t)$ is non-negative for all bounded signed measures $\sigma$, and if non-negativity holds for all $\sigma$ satisfying the condition $\int d\sigma(s) = 0$.)

Definition 1. Given a conditionally non-negative definite $K_G(s, t)$, possibly depending on $G$, the $K$-based quadratic distance between two probability measures $F_\tau$ and $G$ is defined as

$$d_K(F_\tau, G) = \int \int K_G(s, t) \, d(F_\tau - G)(s) \, d(F_\tau - G)(t).$$

Note that by allowing $K$ to depend on $G$ we no longer have an inner product space, but we retain non-negativity owing to conditionally non-negative definiteness.

2.1.2. Estimation of quadratic distance
We now focus on the empirical estimation of the quadratic distance. We shall later use these results for the estimation of quadratic risk. It will be useful to express our results in terms of the $n \times n$ dimensional empirical kernel matrix $\mathbf{K}$ of a data set $x_1, \ldots, x_n$, which we define to have $ij$th
element $K_{ij} = K(x_i, x_j)$. Crucial to the estimation is the concept of the centred kernel, which is defined as follows.

**Definition 2.** The $G$-centred kernel $K$, which is denoted by $K_{cen(G)}$, is defined as

$$ K_{cen(G)}(x, y) = K(x, y) - K(x, G) - K(G, y) + K(G, G), $$

where $K(x, G) = \int K(x, y) \, dG(y)$ and $K(G, G) = \int \int K(x, y) \, dG(x) \, dG(y)$.

For example, the $G$-centred kernel for the Pearson $X^2$-kernel in equation (3) simplifies to

$$ K_{cen(G)}(x, y) = \sum_{i=1}^C \sum_{j=1}^C \left[ I(x \in A_i) \right] \left[ I(y \in A_j) \right] - 1. $$

Using the $G$-centred kernel we can rewrite the distance in the form of a $U$-functional as

$$ d_K(F_\tau, G) = \int \int K_{cen(G)}(x, y) \, dF_\tau(x) \, dF_\tau(y). $$

But as $F_\tau$ is unknown we shall use the empirical cumulative density function $\hat{F}$ to estimate $d_K(F_\tau, G)$. Using equation (7), for any fixed $G$, we arrive at the following two estimates for $d_K(F_\tau, G)$. First, $d_K(\hat{F}, G) = K_{cen(G)}(\hat{F}, \hat{F}) = V_n$ is a $V$-statistic (Serfling, 1980), which is known to provide a biased estimate for the $U$-functional in equation (7). It can be calculated in matrix form as $1^T \mathbb{K}_{cen(G)} 1 / n^2$. We can also construct an unbiased estimate by using the corresponding $U$-statistic:

$$ U_n = \frac{1}{n(n-1)} \sum_i \sum_{j \neq i} K_{cen(G)}(x_i, x_j) = \frac{1}{n(n-1)} \{ 1^T \mathbb{K}_{cen(G)} 1 - tr(\mathbb{K}_{cen(G)}) \}, $$

where $tr$ denotes the usual matrix trace. We shall later use the notation $\text{trace}_G(K)$ to refer to a functional version of the trace operation that is defined by

$$ \text{trace}_G(K) = \int K(x, x) \, dG(x). $$

Also note that, for the Pearson $X^2$-distance, using the estimator $V_n$ we obtain

$$ \sum_{i=1}^C \left\{ \hat{F}(A_i) - G(A_i) \right\}^2 / G(A_i), $$

which equals the Pearson $X^2$-statistic divided by $n$, whereas the unbiased estimator by using equation (8) gives

$$ \sum_{i=1}^C \left\{ \hat{F}(A_i) - G(A_i) \right\}^2 / G(A_i) - (C - 1)/n. $$

Under the usual assumptions, $n$ times expression (10) follows a $\chi^2_{C-1}$-distribution, so its expectation is $C - 1$, whereas the unbiased estimator in expression (11) has expectation 0.

### 2.2. Quadratic risk definition and unbiased estimation

On the basis of the quadratic distance we have already defined the quadratic risk of a parametric model $\mathcal{M}$ as

$$ \rho_d(F_\tau, \mathcal{M}, m) = E\{d_K(F_\tau, M_\theta)\}. $$
Here \( \hat{\theta} \) is an estimator of \( \theta \), based on a sample of size \( m \). In our subsequent analyses we shall assume that \( \hat{\theta} \) is the maximum likelihood estimator of \( \theta \). Note also that \( M_j \) is the second argument in \( d_K(\cdot, \cdot) \), so when we use a \( G \)-dependent kernel (see equation (4)) it will be playing the role of \( G \).

Next we construct an unbiased estimator of the risk that is given in equation (12). We start by showing how to estimate the quadratic risk of the empirical distribution function unbiasedly, a non-parametric estimator of \( F_\tau \). As mentioned earlier, a comparison of the non-parametric risk and the parametric risk gives a global assessment of model quality.

### 2.2.1. Unbiased estimation of empirical risk

For simplicity of notation, we define the risk of the non-parametric fit, based on \( m \) observations, by \( R_m \). A straightforward calculation shows that

\[
R_m = E\{d_K(F_\tau, \hat{F})\} = \frac{1}{m} \text{trace}_F(K_{\text{cen}}(F_\tau)).
\]  

For example, for the Pearson \( X^2 \)-example the above risk can be calculated as

\[
\frac{1}{m} \sum_{i=1}^{C} \left\{ \frac{1}{F_\tau(A_i)} - 1 \right\} F_\tau(A_i) = \frac{1}{m} \sum_{i=1}^{C} \{1 - F_\tau(A_i)\} = \frac{C - 1}{m}.
\]

For most other kernels we shall need to estimate \( R_m \) because it will depend on \( F_\tau \). Provided that \( K \) does not depend on \( G \), we can compute an unbiased estimator of \( \text{trace}_F(K_{\text{cen}}(F_\tau)) \), based on a sample of size \( n \) by using

\[
\frac{1}{n} \sum_{i=1}^{n} K(x_i, x_i) - \frac{1}{n(n-1)} \sum_i \sum_{j \neq i} K(x_i, x_j).
\]

This immediately gives us the unbiased estimator of \( R_m \) as

\[
\hat{R}_m = \frac{1}{m} \left\{ \frac{1}{n} \sum_{i=1}^{n} K(x_i, x_i) - \frac{1}{n(n-1)} \sum_i \sum_{j \neq i} K(x_i, x_j) \right\}.
\]  

Note that

\[
\hat{R}_m \xrightarrow{\text{prob}} 0
\]

as \( n \to \infty \) and, if we set \( m = n \), then \( \hat{R}_n = O_p(1/n) \). Hence any method of systematically selecting models with smaller estimated risk than \( \hat{R}_n \) will give risk estimates converging to 0.

### 2.2.2. Unbiased estimation of risk of parametric models

We next build an unbiased estimator of the risk of the parametric model \( \mathcal{M} = \{M_\theta\} \), which is given in equation (12). This risk can be estimated unbiasedly at any sample size \( m \leq n - 2 \) as follows. Let \( A_m \) be a randomly selected subset of size \( m \) from \( \{1, 2, \ldots, n\} \), and let the point estimator \( \hat{\theta}(A_m) \) be the value of \( \hat{\theta} \) based on \( \{x_i: i \in A_m\} \). Also, we define

\[
U(A_m) = \frac{1}{(n-m)(n-m-1)} \sum_{i,j \in A_m, i \neq j} K_{\text{cen}}(M_{\hat{\theta}(A_m)})(x_i, x_j),
\]  

\( A^c \) being the complement of set \( A \). As \( \hat{\theta} \) is constructed from \( m \) independent observations, \( U(A_m) \) is an unbiased estimator of \( \rho(\mathcal{M}, m) \), for \( m \leq n - 2 \). Using equation (16) we can define
an unbiased estimator by averaging over all possible subsets of size $m$,

$$\hat{\rho}(F_\tau, M, m) = \frac{1}{n \choose m} \sum_{i=1}^{n \choose m} U(A_{m_i}).$$

If we wish to estimate $\rho(F, M, n)$, then $m = n - 2$ might be expected to generate the least bias. However, as we argue at the end of this section, other choices of $m$ might be made for either consistency or parsimony. Also, for computational reasons we might wish to use only a few selected subsets of size $m$, as in $V$-fold cross-validation (van der Laan et al., 2004), or randomly selected subsets of size $m$ (Blom, 1976).

Note that we can construct unbiased estimators of the AIC relative risk in a similar fashion, i.e. we can estimate the first term on the right-hand side of equation (1) as $\Sigma \log \{m(\hat{\theta}_{(-i)})(x_i)\}$. This estimator is unbiased for risk at sample size $m = n - 1$.

### 2.3. Decomposition of quadratic risk and approximate quadratic risk estimators

In many model selection problems, we may choose to simplify the unbiased estimates of risk, which can be both difficult and expensive to compute. The main difficulty is that the point estimators are traditionally obtained by using the notoriously slow EM algorithm and therefore it is difficult to obtain them to sufficiently high precision in each of the many ‘leave out observations’ recalculations. For this reason, we derive an alternative estimate of the quadratic risk, based on asymptotic expansions, similar to those of the AIC derivation.

We start with the following decomposition of the distance:

$$d_K(F_\tau, M_\hat{\theta}) = d_K(F_\tau, M_{\theta_\tau}) + \{d_K(F_\tau, M_\hat{\theta}) - d_K(F_\tau, M_{\theta_\tau})\}.$$  \hspace{1cm} (17)

Here $\theta_\tau$ denotes the parameters of the distribution (in the class of distributions that are denoted by $M_\theta$) closest in quadratic distance to the true model $F_\tau$, i.e. $\theta_\tau = \arg \min_\theta \{d_K(F_\tau, M_\theta)\}$. We shall call the first term on the right-hand side of equation (17) the model lack of fit MLF because it assesses the distance between the best model element and the truth. In particular, it is 0 if $F_\tau \in M_\theta$. Note that MLF does not depend on parameter estimation.

We shall call the second term on the right-hand side of equation (17), $\{d_K(F_\tau, M_\hat{\theta}) - d_K(F_\tau, M_{\theta_\tau})\}$, the parameter estimation error because it measures the deviation of $\hat{\theta}$ from the best parameter $\theta_\tau$. We observe that it is always non-negative, by definition of $\theta_\tau$, and its magnitude increases as $\hat{\theta}$ deviates from $\theta_\tau$. In fact, it is shown in Appendix A (equation (36)) that this term is approximately a quadratic form in $\hat{\theta} - \theta_\tau$, which implies that the parameter estimation cost is a monotonic function of $\hat{\theta} - \theta_\tau$. On the basis of the interpretation of the decomposition of the distance we can rewrite equation (17) as

$$d_K(F_\tau, M_\hat{\theta}) = \text{MLF} + \text{parameter estimation error}.$$  \hspace{1cm} (18)

For the risk calculation we take expectations in equation (18) to obtain

$$E\{d_K(F_\tau, M_\hat{\theta})\} = \text{MLF} + \text{PEC}_{(m)},$$  \hspace{1cm} (19)

where $\text{PEC}_{(m)} = E\{d_K(F_\tau, M_\hat{\theta}) - d_K(F_\tau, M_{\theta_\tau})\}$ represents the parameter estimation cost when $\hat{\theta}$ comes from a sample of size $m$.

This decomposition is essential both for finding an approximation to the overall risk and for providing a useful interpretation of the different errors driving the risk. The approximation and the estimates for the two terms, MLF and PEC, will be directly based on the asymptotic theory of quadratic distance (Lindsay et al., 2007) and the unbiased estimators of quadratic distance.
Now we shall state a few results from Lindsay et al. (2007) and use these results to justify our approximation and finally provide an appropriate estimator for the quadratic risk.

2.3.1. Kernel projection operator and score-centred kernel
Crucial to these approximations of quadratic risk are the concept of the scored centred kernel and the score-based projection operators which we now define.

For a parametric model given by \( G_\theta \), with density given by \( g(x; \theta) \), let us denote the set of score functions by \( \nabla \log \{g(x_i; \theta)\} = s(x_i; \theta) \), where \( \nabla \) is the vector differential with respect to every element of \( \theta \). Note that the maximum likelihood estimator \( \hat{\theta} \) is obtained as a solution to \( \Sigma s(x_i; \theta) = 0 \). Further, we define the extended score vector \( u_\theta^*(x) = (1, s(x, \theta)^T)^T \) and the extended information matrix for a single observation to be

\[
J_\theta^* = E_\theta(u_\theta^*u_\theta^*^T).
\]

We shall then let \( P^* \) be the kernel operator that is defined by

\[
P^*_\theta(x, y) = u_\theta^*(x)^T J_\theta^*^{-1} u_\theta^*(y),
\]

i.e. \( P^*_\theta \) acts as a projection operator on the likelihood scores \( u_\theta^* \). The score-centred kernel \( K_{scen}(G_\theta)(x, y) \), as centred under \( G_\theta \), is defined to be

\[
K_{scen}(G_\theta)(x, y) = (I - P^*_\theta)K(I - P^*_\theta)
= K(x, y) - \int P^*_\theta(x, z) K(z, y) \, dG_\theta(z) - \int K(x, z) P^*_\theta(z, y) \, dG_\theta(z)
+ \int \int P^*_\theta(x, z) K(z, w) P^*_\theta(w, y) \, dG_\theta(z) \, dG_\theta(w).
\]

Similar to the matrix form of \( \mathbb{K} \), we define the \( n \times n \) matrix \( P_\theta = u_\theta^* J_\theta^{-1} u_\theta^T \), where \( u_\theta^* \) is the \( n \times p \) matrix with entries \( \partial_{\theta_j} \log \{g_\theta(x_i)\} \), \( x_i \) being the \( r \)th data point and \( \theta_j \) being the \( j \)th component of \( \theta \). Similarly we represent the matrix version of the scored centred kernel as \( K_{scen}(\theta) \), which can be calculated as

\[
K_{scen}(\theta) = (I - P^*_\theta)K(I - P^*_\theta)
= (I - P^*_\theta)K_{scen}(\theta) = (I - P^*_\theta)K_{scen}(\theta).
\]

The first important property of score centring is the following alternative representation for the empirical distance between the data and the estimated model:

\[
d_K(\hat{F}, G_\hat{\theta}) = \int \int K_{scen}(\hat{G}_\theta)(x, y) \, d\hat{F}(x) \, d\hat{F}(y),
\]

which has the following asymptotic property.

Theorem 1 (Lindsay et al., 2007). Given the regularity assumptions that are stated in Lindsay et al. (2007), under \( G_\hat{\theta} \) we have

\[
n \left\{ d_K(\hat{F}, G_\hat{\theta}) - \int \int K_{scen}(\hat{G}_\theta)(x, y) \, d\hat{F}(x) \, d\hat{F}(y) \right\} \overset{\text{prob}}{\to} 0.
\]

This result can then be used to show that \( n \, d_K(\hat{F}, G_\theta) \) has an asymptotic distribution that can be represented as \( \Sigma_{i=1}^\infty \lambda_i \chi_i^2 \), where the \( \lambda_i \)'s are the eigenvalues of the spectral decomposition of kernel \( K_{scen}(G_\theta) \). The asymptotic mean is therefore \( \Sigma_{i=1}^\infty \lambda_i = \text{trace}(K_{scen}(G_\theta)) \) (see Lindsay et al. (2007) for details).
2.3.2. Approximate quadratic risk estimators

We now return to our original purpose of approximating the different terms of our quadratic risk. These approximations will be based on asymptotic calculations for each model, assuming that the model is correct. This same technique is used in the AIC calculation, where it clearly produces great simplification in risk estimation. We start with the asymptotic approximation to PEC.

**Proposition 1.** Suppose that $M_\theta$ has the density function $m_\theta$ and that the true model $F_\tau$ is in the model space $M_\theta$, having parameter $\theta_\tau$. In this case $d_K(F_\tau, M_{\hat{\theta}}) = 0$. Define

$$V(x, y) = \int \int P(x, z) K(z, w) P(w, y) M_{\theta_\tau}(z) M_{\theta_\tau}(w)$$

where $P(x, y) = u(x)^T J^{-1} u(y)$. Then as $m$, the sample size for $\hat{\theta}$, goes to $\infty$

$$m d_K(F_\tau, M_\theta) - m \int \int V(x, y) \hat{F}(x) \hat{F}(y) \xrightarrow{\text{prob}} 0.$$ 

It follows that the asymptotic mean of $m d_K(F_\tau, M_\theta)$ is

$$\text{tr}_{\theta_\tau} V = \int \int V(x, y) M_{\theta_\tau}(x) M_{\theta_\tau}(y).$$

For a proof, see Appendix A.

As a simple example, for the Pearson $X^2$-kernel PEC is simply $\text{dim}(\hat{\theta})/m$, where $\text{dim}(\hat{\theta})$ denotes the number of parameters being estimated. In general we must estimate PEC and, for an estimator of trace $(P_\theta^* K P_\theta^*)$, we use its empirical version $(1/n) \text{tr}(P_\theta^* K P_\theta^*)$ based on the full sample. This yields the estimator

$$\hat{\text{PEC}}_m = \frac{1}{m n} \text{tr}(P_\theta^* K P_\theta^*).$$

(23)

Now we turn to the estimation of MLF = $d_K(F_\tau, M_{\theta_\tau})$, which is a term that does not depend on hypothetical sample size $m$. We start with its sample equivalent given by

$$d_K(\hat{F}, M_\theta) = \int \int K_{\text{scen}}(M_{\theta_\tau})(x, y) d\hat{F}(x) d\hat{F}(y) = \frac{1}{n^2} I^T K_{\text{scen}}(M_{\hat{\theta}}) I. \quad (24)$$

This estimator has bias $E\{d_K(\hat{F}, M_\theta)\} - \text{MLF}$. To do bias correction we again use an asymptotic approximation assuming that $F_\tau = M_{\theta_\tau}$, so that MLF = 0. Applying the remarks following theorem 1 we obtain the following approximation to the bias term:

$$E\{d_K(\hat{F}, M_\theta)\} - \text{MLF} \approx \frac{1}{n} \text{tr} \{K_{\text{scen}}(M_{\theta_\tau})\}. \quad (25)$$

The right-hand side of approximation (25) can be estimated by its sample equivalent

$$\frac{1}{n^2} \text{tr}(I - P_\theta^*) K(I - P_\theta^*). \quad (26)$$

Thus, we have the bias-corrected estimator of MLF,

$$\hat{\text{MLF}} = d_k(\hat{F}, M_\theta) - \frac{1}{n^2} \text{tr}(I - P_\theta^*) K(I - P_\theta^*). \quad (27)$$
For example, in the Pearson $X^2$-case expression (26) reduces to \( \{C - 1 - \dim(\hat{\theta})\}/n \) and thus MLF can be estimated as

\[
\sum_{i=1}^{C} \frac{\{\hat{F}(A_i) - G_{\hat{\theta}}(A_i)\}^2}{G_{\hat{\theta}}(A_i)} - \frac{C - 1 - \dim(\hat{\theta})}{n}.
\]

Since \( C - 1 - \dim(\hat{\theta}) \) gives the residual degrees of freedom for the $C$-cell Pearson $X^2$-test it is clear how the correction removes the bias that is inherent in the $\chi^2$ limiting distribution.

Finally combining the estimators of PEC$_{(m)}$ and MLF that are given in equations (23) and (27) we have the following estimator for the quadratic risk:

\[
\rho_{dk}(\mathcal{M}, m) = \frac{1}{n^2} \text{tr}(\mathbb{I} - \mathbb{P}^*_{\hat{\theta}})\mathbb{K}(\mathbb{I} - \mathbb{P}^*_{\hat{\theta}}) + \frac{1}{nm} \text{tr}(\mathbb{P}^*_{\hat{\theta}} \mathbb{K} \mathbb{P}^*_{\hat{\theta}}) = \frac{1}{n^2} \mathbb{I}^T \mathbb{K}_{\text{cen}}(\hat{\theta}) \mathbb{I} - \frac{1}{n^2} \text{tr}(\mathbb{I} - \mathbb{P}^*_{\hat{\theta}})\mathbb{K}(\mathbb{I} - \mathbb{P}^*_{\hat{\theta}}) + \frac{1}{nm} \text{tr}(\mathbb{P}^*_{\hat{\theta}} \mathbb{K} \mathbb{P}^*_{\hat{\theta}}).
\]

Note that for the Pearson $X^2$-example the risk estimator, when calculated at $m = n$, simply becomes

\[
\sum_{i=1}^{C} \frac{\{\hat{F}(A_i) - G_{\hat{\theta}}(A_i)\}^2}{G_{\hat{\theta}}(A_i)} - \frac{\{C - 1 + 2 \dim(\hat{\theta})\}}{n}.
\]

Observe that the risk estimator that is given by equation (28) is numerically less expensive than the cross-validation-based exact estimator that is given in equation (16). Unlike the unbiased risk estimator, it does not require repeated parameter estimation. Moreover, the matrix operations that are involved in the calculation are also inexpensive, making equation (28) a promising risk estimator.

For model comparison purposes we shall also want to estimate the risk of the empirical distribution function $\hat{F}$, in which case $\mathbb{P}^*_{\hat{\theta}} = \mathbb{I}$. The estimated distance $d_K(\hat{F}, \hat{F})$ is 0. So a biased estimator of $R_m$, the risk of the $\hat{F}$, at sample size $m$, may be calculated by using

\[
\hat{R}^{\text{b}}_m = \frac{1}{nm} \text{tr}(\mathbb{K}_{\text{cen}}(\hat{F})).
\]

We shall later see that, if we use an appropriately scaled kernel $K^s$ ($K^s = K/c$, where $c = \text{tr}(K)/\text{tr}(K^2)$), $\hat{R}^{\text{b}}_m$ is exactly equal to the degrees of freedom corresponding to the scaled kernel $K^s$. In fact for the Pearson $X^2$-kernel $\hat{R}^{\text{b}}_m = C - 1$, which is the degrees of freedom for a $C$-cell $\chi^2$ goodness-of-fit test.

In the special case that the estimated risk is evaluated at $m = n$ we shall call it the quadratic AIC risk QAIc, where AIC reflects commonality with the AIC derivation. We shall now show how to mimic the BIC by using a different value of $m$. Recall that, for the usual AIC with KL loss function, the relative MLF, after subtracting $\int \log \{f_{z}(x)\} f_{z}(x) \, dx$ common to all models corresponds to $-\int \log \{m_{\hat{\theta}_{\text{L}}}\} f_{z}(x)$ (see equation (1)), which has the approximately unbiased estimator (under the model):

\[
\widehat{\text{MLF}} = -2 \frac{\hat{l}}{n} + \frac{\dim(\hat{\theta})}{n}.
\]

The parameter estimation cost, $\widehat{\text{PEC}}_{(m)} = \dim(\hat{\theta})/m$, giving us the estimated risk,

\[
-2 \frac{\hat{l}}{n} + \frac{\dim(\hat{\theta})}{n} + \frac{\dim(\hat{\theta})}{m}.
\]
For AIC risk we use $\text{PEC}(n)$, which gives us $\text{AIC}/n = -2\hat{l}/n + 2 \dim(\hat{\theta})/n$. Note the similarity of this risk to the Pearson risk estimator in expression (29). The essential difference is that the latter estimates absolute risk, not relative risk. If we use $m=n/\{\log(n) - 1\}$ in expression (31), we arrive at the standard BIC formula

$$\frac{\text{BIC}}{n} = -2\frac{\hat{l}}{n} + \frac{\log(n) \dim(\hat{\theta})}{n}.$$ 

For this reason the estimated risk (28) evaluated at $m=n/\{\log(n) - 1\}$ will be called QBIC and can be written as

$$\text{QBIC} = \hat{\text{MLF}} + \{\log(n) - 1\} \hat{\text{PEC}}(n).$$

Along with a recipe for estimation of the risk, the decomposition in expression (28) provides an excellent tool for deeper understanding and analysis of the interplay of model misspecification and the effective parameter cost of using the model. This provides us with the ingredients for constructing alternative model selection tools and tools for accessing global fits which we discuss in the following section.

### 2.4. Consistency and choice of $m$

It seems very natural to assess the risk of a model at the sample size $m=n$. This risk indicates how well the chosen model, along with the current parameter estimates, might be expected to perform when used to approximate the true distribution. Unfortunately, estimating the risk at $m=n$ is a difficult problem: we cannot estimate this risk sufficiently well to discriminate consistently between two true models with different numbers of parameters.

In this section we shall consider the meaning of consistency under several asymptotic scenarios. We shall use the AIC method to illustrate our points because of its simple form. In all of them, we shall suppose that the true distribution is in one of the models, i.e. $F_\tau \in M$.

First, if we leave $m$ fixed as $n \to \infty$, then the best model, in terms of risk, need not contain the true distribution, and consistency would mean selection of this (possibly false) best model. This might be desirable for reasons of parsimony, as the true models may be too complex to be useful.

In a formal sense, methods like the AIC cannot be consistent in this asymptotic scenario because they use asymptotic approximations for the cost of estimating parameters, and those approximations assume that the model is true. If we replace the exact risk with the asymptotic approximation, however, consistency is achieved. See, for example, the AIC in expression (31), where the $\hat{\text{MLF}}$-term is consistent for MLF, and the approximate parameter estimation cost is just a function of parameter dimension when $m$ is fixed.

A second way to view consistency is to let $m=m/n$ increase with $n$. In this case, the parameter estimation costs shrink to zero, and so consistency would involve selecting a true model. A stringent form of consistency, but natural, would require the method to select the smallest true model with probability approaching 1. We shall call this smallest true model consistency.

It is well known that AIC, with $m_n=n$, lacks this form of consistency. When two true models $M_1 \subset M_2$ are nested (and regularity conditions hold), the probability that the AIC picks the larger of the two models corresponds to using a likelihood ratio test with the constant critical value $2\{\dim(M_2) - \dim(M_1)\}$. For example if the difference of dimension is 1, with probability $0.16 (=\Pr(\chi^2_1 > 2))$ we choose the larger model.
It turns out that smallest true model consistency holds when \( m/n \to 0 \). The proportional case, when \( m \approx \gamma n \), for constant \( \gamma > 0 \), is at the cutting edge of consistency. It is easily checked that, for the AIC, it is equivalent to using the critical value \( (1 + \gamma^{-1}) \{ \dim(M_2) - \dim(M_1) \} \).

For example, if the difference in dimension is 1, the probability of overshooting is around 0.045, 0.025, 0.01 and 0.001 for the proportion \( \gamma = 1/3, 1/4, 1/5, 1/10 \) respectively. Thus, as \( \gamma \) approaches 0, the probability of overshooting goes to 0.

Although we might choose to use a BIC-type criterion for its familiarity, we think that it would be very useful also to describe the null probability of rejection for the corresponding likelihood ratio critical value.

3. Model selection by using quadratic risk

Now, we provide a short description of how we might use quadratic risk as a model selection tool. First, we need to understand the interplay of the terms in the risk decomposition. See Fig. 1 for an illustration. It shows the relevant quantities for an example in model (3) with \( n = 300 \), which is described later in Section 5. Suppose that we have a sequence of models \( M_k \) that are nested with increasing \( k \). As model complexity \( k \) increases, the empirical distance \( d_K(\hat{F}, \hat{M}_k) \) decreases but must stay non-negative. We also know that MLF, for any model that contains \( F_\tau \), will tend to 0 in probability. The estimated parameter estimation cost PEC starts near 0 and increases with model complexity. However, it never will be larger than the risk of the empirical estimate, which is the parameter estimation cost of a non-parametric fit. By construction, the total quadratic risk is the simple addition of MLF and PEC. Thus models with low estimated risk arise as a compromise between the decreasing MLF and the increasing PEC. A simple model selection rule is to pick the model having the minimum value of QAIC.

The QBIC form of the risk, which is given in equation (32), introduces a larger penalty for parameter estimation than QAIC. In Fig. 1(b) we can see how this alteration has created a much sharper minimum in the risk function. Both for QAIC and QBIC, we shall say that the model with the minimum estimated risk is the best model.

Just like the AIC and BIC, quadratic risks need not attain an absolute minimum in the range of models that is considered. However, there is an important benchmark, given by the empirical risk, for evaluating the performance of a model when using quadratic risk measures. Note that for the empirical distribution there is no model lack of fit, because the model is non-parametric, but there is the maximal possible parameter estimation cost. Models that do not meet this benchmark clearly suffer from substantial model lack of fit and so we shall label them as ‘inadequate models’, whereas the rest will be called ‘adequate’. Even in the presence of a minimum risk model, if all our proposed models are ‘inadequate’ it indicates that more model building in the same class or exploration in a larger class of models might be necessary. For example, on the basis of the QAIC risk for the models proposed and empirical model in Fig. 1(b), if we had explored only the five smallest models, then QAIC and QBIC would have selected model 3. However, on the basis of the MRA criterion no model would have been found adequate, thus providing the crucial global framework for model selection.

However, if we have more than one adequate models we shall call the smallest model in this set the minimal risk adequate (MRA) model. Note that the consistency of the empirical risk (which is shown in Section 2.2.1) ensures that this model selection method would have estimated risk going to 0 as \( n \to \infty \). Now, inspecting Fig. 1(b) for the whole range of 10 components model, we see that all models from 6 onwards are adequate. Thus model 6 is the MRA model, even though the QAIC selects model 9 as the minimum risk model.

We shall provide illustrations of these criteria in the next section.
Fig. 1. Illustration of (a) components of the risk measure (–■–, distance; ●●●●, MLF; ●●●●, PEC) and (b) the risk measures (–■–, QAIC; ●●●●, empirical QAIC risk; – – –, QBIC) corresponding to a single run of the simulation of model 3 that is described in Section 5.1.1.
4. Application: selecting the number of components in mixture models

Now we apply our quadratic risk model selection methodology to the problem of selecting the number of components in a multivariate normal mixture model. We start by introducing the following notation and definitions for mixture models.

A random variable $X \in \mathbb{R}^D$ is said to follow a $k$-component normal mixture model if its density $f^{(k)}$ can be written as

$$f^{(k)}(x; \mu, \Sigma, \pi) = \sum_{j=1}^{k} \pi_j \phi(x; \mu_j, \Sigma_j)$$

for $x \in \mathbb{R}^D$, \(33\)

where $\phi(x; \mu_j, \Sigma_j)$ denotes a normal density with mean $\mu_j$ and variance $\Sigma_j$, and $\pi_j < 1 \, \forall j$ and $\Sigma_j = 1$. For compactness we denote $\theta = \{\mu, \Sigma, \pi\}$, with the above restrictions, and denote the density in expression (33) as $f^{(k)}_\theta$.

According to the general framework, now $\mathcal{M}$ denotes the class of all $D$-dimensional finite mixtures with normal mixing components, i.e. for a fixed $D$ our $\mathcal{M} = \{M_1, M_2, \ldots \}$, where $M_k$ denotes the $k$-component normal mixture models, and for a particular model element $M_\theta \in \mathcal{M}_k$, which is more precisely denoted as $M^{(k)}_\theta$, we have $dM^{(k)}_\theta = f^{(k)}_\theta$.

The complexity of the null distribution makes a testing theory for the number of components rather more difficult. It also means that the standard regularity assumptions behind the derivations of the AIC and BIC are false. For a description of the previous approaches to selecting the number of components of mixture models see McLachlan and Peel (2000).

4.1. Risk-based analysis of mixture complexity

We now develop a strategy for selecting the number of components by using our quadratic risk assessment approach. First, we shall specify the kernel to specify the loss function. We shall then outline the steps for estimating the risk function and finally illustrate the use of the risk measure in choosing a mixture model.

4.1.1. Specifying the kernel

In principle, any non-negative definite kernel can be used to build the distance measure. However, from the results in Section 2 it is clear that the key to the calculation of the distance, and hence the risk, are the integrals that are required in the definition of the quadratic distance. These will involve high dimensional numerical integration for arbitrary choice of the kernel $K$. For this reason it is desirable to use kernels for which the integrals $\int K(x,y) \, dM(y)$ and $\int \int K(x,y) \, dM(x) \, dM(y)$ can be carried out explicitly.

When the model is a finite mixture of normals, the multivariate normal kernel meets this goal. The $D$-dimensional normal kernel, in its most general form, is defined as

$$K_\Sigma(x, y) = \frac{1}{(2\pi)^{D/2} |\Sigma|^{1/2}} \exp\left\{ -\frac{1}{2} (x - y)^T \Sigma^{-1} (x - y) \right\}.$$ 

We shall take $\Sigma = h^2 I$, $h$ being a ‘smoothing parameter’, and henceforth denote the corresponding kernel as $K^h$. This constant $h$ can be thought of as a ‘tuning’ parameter, which is something analogous to the bin width in the construction of a histogram. We shall discuss its role further in the next subsection.

4.1.2. Role of tuning parameter $h$ and its empirical estimation

Lindsay et al. (2007) developed a tool for the understanding of the operating characteristics of a quadratic distance that they called the spectral degrees of freedom of the kernel, sDOF. It
is a generalization of the degrees of freedom of the $\chi^2$-distance (itself a quadratic distance) to other kernels by examining their limiting distributions. With this tool we can roughly equate the degree of smoothing that comes from a choice of $h$ to that of the $\chi^2$-test having the same degrees of freedom, which in turn corresponds to the number of cells that are used in its construction. The tuning parameter $h$ is analogous to choosing the bin width of each cell (or, equivalently, choosing the number of cells) in a $\chi^2$ goodness-of-fit test.

The spectral degrees of freedom of the kernel depend on the true sampling distribution and the kernel $K$ (whether centred or score centred) through the formula

$$sDOF = \left\{ \int K^h(x,x) \, dF_\tau(x) \right\}^2 \left/ \int K^h(x,x)^2 \, dF_\tau(x) \right. .$$

In this paper, we shall base our selection of $h$ on the empirical estimate of $sDOF$ that is given by

$$\hat{sDOF} = \frac{\text{tr}(\hat{K}^h_{\text{cen}}(\hat{F}))^2}{\text{tr}\{\hat{K}^h_{\text{cen}}(\hat{F})\}^2},$$

where the centred kernel matrix based on the empirical distribution has the $ij$th element

$$K^h_{\text{cen}}(x_i,x_j) = K^h(x_i,x_j) - \frac{1}{n} \sum_i K^h(x_i,x_j) - \frac{1}{n} \sum_j K^h(x_i,x_j) + \frac{1}{n^2} \sum_i \sum_j K^h(x_i,x_j).$$

Just as in a $\chi^2$ goodness-of-fit test, there are choices for the spectral degrees of freedom that are clearly too small and others that are clearly too large. As a rough rule of thumb for selecting the number of cells for $D=1$ we suggest that the degrees of freedom should be more than 5 and less than $n/5$, $n$ being the total number of observations. For $D>1$ dimensions we need to increase the lower bound of 5 because the goodness-of-fit test now must assess fit in several directions. We find that, if $sDOF$ is smaller than $(D+1)/2$, it is likely to be too much smoothing, whereas, if $sDOF > n/5$, it would mean too little smoothing. See Section 5 for more on the role of the smoothing parameter on model selection.

4.2. Constructing the risk estimators for mixture model selection

Until now we have focused on the form and the smoothing parameter of the kernel for our problem. Using this kernel we shall now illustrate the steps in estimating the risk of each parametric mixture model $\mathcal{M}_k$, $k=1,2,\ldots$, along with the risk of the empirical model. First, we compute the estimates of the parameters $\hat{\theta}_k$ for each parametric model $\mathcal{M}_k$. Simultaneously, we construct a projection matrix $P^{\hat{\theta}_k}_{\mathcal{M}_k}$ for each model, which is strictly based on the estimates $\hat{\theta}_k$ and the observations. We also calculate the score-centred kernel $\hat{K}^{\hat{\theta}_k}_{\text{scen}(\mathcal{M}_k)}$. Then we calculate the risk estimate of each model by using equations (28) and (32).

5. Simulation results and applications

Now we apply our model selection tool to select the number of components in mixtures. The following steps were used for choosing models in all the examples of this section. First, we standardized the data (i.e. the variance of each variable was scaled to 1). This step was done because we used the same $h$ for all the variables. In this simulation study we restrict ourselves to normal kernels based on the discussion in Section 4.1.1. Using the results in Section 4.1.2 we estimated $sDOF$, which gives us a range of interesting smoothing parameter values. The simulation results that are reported in this section used one representative value from this range, though the results were quite stable across our suggested range. The maximum likelihood estimates for the range of
models under consideration are calculated by using an EM algorithm and the projection matrices $P_{\hat{\theta}}^*$ based on this estimate. Note that for analysis at different levels of smoothing we only need to recalculate the distance and computationally simple matrix $K$, and not the projection matrix $P_{\hat{\theta}}^*$.

5.1. Simulation experiment
Given a mixture parameter set and a fixed sample size, we generated 100 sets of random samples from the true mixture distribution. We have tabulated the frequency of selection of the number of components by each of the five model selection methods (AIC, QAIC, MRA, BIC and QBIC) under different examples, in an experimental design that varies by dimension, sample size and separateness of the components. For some examples we also vary the range of models in the model space.

Fig. 2. Representative data sets from two-dimensional simulation: (a) model 1; (b) model 2; (c) model 3; (d) model 4
Here, we have fixed the variance structure, so our model selection only refers to selecting the number of components. Readers who are familiar with the MCLUST software should note that Fraley and Raftery (1999) have provided a unified strategy to select both the number of components and the variance structure, but only with a single criterion, the BIC. We now discuss the results of our simulation experiment.

### 5.1.1. Two-dimensional simulation

In our first study we held the data dimension to 2 and varied the normal mixture model as well as the sample size (for some cases). The scatterplot of a representative sample from the simulation of each of the four models is given in Fig. 2.

First we consider simulations from the first three models which are generated from two-, four- and six-component normal mixtures respectively. From Table 1 it is clear that AIC and QAIC are the more liberal methods—in the sense of favouring models that are larger than the true model. Although QBIC and BIC are both potentially conservative, QBIC is slightly less so and made model selection errors on both sides of the true value. The MRA criterion (here using the \( m=n \) risk) was fairly conservative in some cases (e.g. simulation 1) but otherwise fairly competitive. Like the BIC it never overshot the true model.

<table>
<thead>
<tr>
<th>Method</th>
<th>Results (%) for the following estimated number of components:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1   2  3  4  5  6  7  8  9  &gt;10</td>
</tr>
<tr>
<td>Model 1: 2 moderately overlapping clusters, 2 dimensions, ( n=1000 )</td>
<td></td>
</tr>
<tr>
<td>AIC</td>
<td>—</td>
</tr>
<tr>
<td>QAIC</td>
<td>5</td>
</tr>
<tr>
<td>MRA</td>
<td>100</td>
</tr>
<tr>
<td>BIC</td>
<td>48</td>
</tr>
<tr>
<td>QBIC</td>
<td>21</td>
</tr>
<tr>
<td>Model 2: 4 distinct clusters, 2 dimensions, ( n=1000 )</td>
<td></td>
</tr>
<tr>
<td>AIC</td>
<td>—</td>
</tr>
<tr>
<td>QAIC</td>
<td>—</td>
</tr>
<tr>
<td>MRA</td>
<td>—</td>
</tr>
<tr>
<td>BIC</td>
<td>—</td>
</tr>
<tr>
<td>QBIC</td>
<td>—</td>
</tr>
<tr>
<td>Model 3: 6 distinct clusters, 2 dimensions, ( n=1000 )</td>
<td></td>
</tr>
<tr>
<td>AIC</td>
<td>—</td>
</tr>
<tr>
<td>QAIC</td>
<td>—</td>
</tr>
<tr>
<td>MRA</td>
<td>—</td>
</tr>
<tr>
<td>BIC</td>
<td>—</td>
</tr>
<tr>
<td>QBIC</td>
<td>—</td>
</tr>
<tr>
<td>Model 3: 6 distinct clusters, 2 dimensions, ( n=300 )</td>
<td></td>
</tr>
<tr>
<td>AIC</td>
<td>—</td>
</tr>
<tr>
<td>QAIC</td>
<td>—</td>
</tr>
<tr>
<td>MRA</td>
<td>—</td>
</tr>
<tr>
<td>BIC</td>
<td>—</td>
</tr>
<tr>
<td>QBIC</td>
<td>—</td>
</tr>
</tbody>
</table>

†Selection of the true model.
Next we explore the effect of sample size on model selection in model 3. Decreasing the sample size from 1000 to 300 tended to increase the underestimation of the three conservative methods but did not greatly degrade the liberal estimates of the AIC and QAIC, reflecting their lack of consistency.

Overall QAIC performed much better than the AIC. Though QAIC overestimated the complexity, in comparison with the AIC the distribution of the number of components was much more peaked at the true number whereas the right-hand tail tapered off very quickly. Also, if the components had high degree of overlap, in the sense of not displaying distinct modes for distinct components, then QBIC performed better than the BIC, which most often underestimated the number of components.

Note also that the quadratic-risk-based methods are more stable with a change of sample size. In those cases where the sample size was small and BIC badly underestimated, the QBIC and MRA methods were distinctly better. Finally, the simulation study revealed that adequacy is a very useful measure. In many examples it provided extra information about the minimal number of components that would be needed to explain the data in hand. Often the minimal adequate model had a considerably smaller number of components than the number with minimum risk QAIC model.

5.2. Quadratic risk as global measure of risk
Next we explore two examples where the true model is beyond our model space and show how the quadratic risk measure provides an excellent global model selection criterion in identifying the scenario. We have already observed that, with \( n = 1000 \) in model 3 with six distinct normal components, QAIC, QBIC, MRA and BIC all overwhelmingly select six components. But what happens if we explore only up to five components, i.e. we do not include the true model in the model class? Table 2 summarizes the results for the five different criteria when we explore only up to five components. We immediately observe that the BIC selects the three-component model in 100% of the cases. Independently our quadratic measures also selected this model in 52% and 87% of the cases, but the MRA criterion shows that none of the models are adequate. Thus if we based our conclusion on the BIC alone we would have failed to recognize that we are trapped in a local minimum. But if we use the quadratic risk measure along with the MRA criterion we are forced to explore models beyond three components as we have a clear indication that even the five-component model is not adequate.

<table>
<thead>
<tr>
<th>Method</th>
<th>Results (%) for the following estimated numbers of components:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>AIC</td>
<td>—</td>
</tr>
<tr>
<td>QAIC</td>
<td>—</td>
</tr>
<tr>
<td>MRA</td>
<td>—</td>
</tr>
<tr>
<td>BIC</td>
<td>—</td>
</tr>
<tr>
<td>QBIC</td>
<td>—</td>
</tr>
</tbody>
</table>

†Model 3: six components, two dimensions, \( n = 1000 \) (limited to five components).
Table 3. Model selection results for the U-example (model (4))

<table>
<thead>
<tr>
<th>Method</th>
<th>Results (%) for the following estimated numbers of components:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>AIC</td>
<td>—</td>
</tr>
<tr>
<td>QAIC</td>
<td>—</td>
</tr>
<tr>
<td>MRA</td>
<td>—</td>
</tr>
<tr>
<td>BIC</td>
<td>—</td>
</tr>
<tr>
<td>QBIC</td>
<td>—</td>
</tr>
</tbody>
</table>

We now explore an example where the model is not a normal mixture, but rather when the scatter of the distribution has the shape of the letter U, as given in Fig. 2 (model 4). These data sets were simulated by first generating the $x$-co-ordinate uniformly between $-1.5$ and $1.5$ and then generating $y$ uniformly between $x^2 - 1$ and $x^2 + 1$. By design, this density is very difficult to capture with a few normal components. So we explored up to 14 normal components. The BIC and QBIC mostly choose 8–10 components (Table 3). Even using the liberal measure QAIC, 51% of the times we choose a model with fewer than 14 components. But the MRA criterion overwhelmingly (70% of the cases) shows that even a 14-component model is not adequate. This suggests that we should either explore beyond a 14-component model or we should explore beyond normal mixtures.

In both these examples it can be easily seen that the BIC, which is the most widely used model selection method for mixtures, does not provide the best result, simply because it is a relative rather than a global measure of risk. In contrast in each of these simulations the MRA criterion clearly points out that the minimum risk model does not lie in the model class that we are considering.

5.3. High dimensional simulation

We now examine the effect of increasing dimensions. Our four-, eight- and 12-dimensional examples with six components were generated in the following way: the first two dimensions are the same as the two-dimensional model 3 that was used in Section 5.1.1; the remaining dimensions are generated from standard normal variates. Like the two-dimensional model 3 the performance of the five criteria are judged under two scenarios:

(a) when the true model is in the model space and
(b) when the true model was not included in the search space.

The results for both scenarios are given in Table 4, the left-hand panel corresponding to scenario (a), where we explore up to 10 components, and the right-hand panel corresponding to scenario (b) where we restrict ourselves to fewer than six components.

First we summarize the results from scenario (a). The AIC grossly overestimates the number of components. QAIC overestimated the number of components but less so than the AIC, but more importantly its performance did not deteriorate much with the increase in dimensions. As dimension increases, BIC tends to underestimate the true number of components with higher percentage, and in many cases (20%, 23% and 25% for four, eight and 12 dimensions respectively) the BIC even chose the three-component model. QBIC and MRA erred on either side of the true model, but both of them chose the true model with a higher percentage than the BIC.
Table 4. High dimensional simulation results

<table>
<thead>
<tr>
<th>Method</th>
<th>Results (%) for the following estimated numbers of components:</th>
<th>Results (%) for the following estimated numbers of components (search limited to 5 components):</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Model 5: 6 components, n = 1000 in 4 dimensions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AIC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>QAIC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MRA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BIC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>QBIC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model 6: 6 components, n = 1000 in 8 dimensions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AIC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>QAIC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MRA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BIC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>QBIC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model 7: 6 components, n = 1000 in 12 dimensions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AIC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>QAIC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MRA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BIC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>QBIC</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

†Selection of the true model.

Moreover, when QBIC and MRA underestimated they erred slightly selecting five components, whereas the BIC selected the three-component model in many cases.

For scenario (b) the results are consistent with the findings of the two-dimensional case, i.e. the various information criteria overwhelmingly suggest that the three-component model is the true model, whereas the MRA criterion combined with the quadratic measures clearly points out (in 96%, 92% and 90% for four, eight and 12 dimensions respectively) that we have not reached a global minimum and we should expand our search space.

These examples clearly show the usefulness of quadratic-risk-based methods in high dimensions, both for selecting the true model and for providing a global infrastructure for model selection.

5.4. Application to real data set

We apply our model selection criterion to a real data set where the goal is to identify the number of gene groups displaying distinctive gene expression profiles under a set of conditions. The details of the experiment are described in Chitikila et al. (2002) and our data set was obtained through personal communication from Dr F. Pugh of Pennsylvania State University. Here we present a very short description of the experimental design and preprocessing of the data.

The experiment was designed to answer an interesting question in transcriptional genomics: is the binding phenomenon of TATA binding proteins (TBPs) affected by neighbouring TAF1 proteins? The TBPs bind to the TATA box and play a crucial role in transcription, which finally results in protein synthesis. The TBP is intertwined to form a three-dimensional structure in such a way that it forms a concave surface and a convex surface; the concave surface attaches to
the TATA box, whereas the convex surface may attach to some other protein, e.g. the TAF N
terminal domain region of TAF1 proteins. To understand this interplay of TBP and the TAF
N terminal domain region of TAF1, a set of 19 experiments were designed using 6226 yeast
genes for each strain. The two main factors that were varied in these experiments are

(a) whether the TAF1 proteins of the yeast cells had the TAF N terminal domain regions
intact (wild types) or whether the TAF N terminal domain regions were removed (Δ-
TANDs) and

(b) whether the concave region of the TBP proteins remained unmutated (wild type) or
whether a yeast strain with mutation was used.

A further three different types of mutations and a few controls were also used. Gene clusters
displaying distinct gene expression profiles, one for each of the four combinations of the two
levels of the two major factors, would suggest an ‘interaction’ model, whereas only two clusters
would suggest that TBPs are not influenced by neighbouring proteins.

The analysis was done as follows. After leaving out genes which did not show significant
change over different conditions, we analysed \( n = 2809 \) genes under \( d = 19 \) conditions (dimen-

---

**Fig. 3.** Risk calculation for the yeast gene expression data set
Model Selection in High Dimensions

117

sions), completely ignoring the information on the identity of the conditions. We used normal mixtures to fit the data and explored up to eight components. The risk curves of each of the four model selection criteria are given in Fig. 3. On the basis of the minimum risk, the BIC chose three components whereas the AIC chose nine components. For calculating the quadratic risk we first applied the sDOF analysis that was detailed in Section 4.2.2 and chose a range of $h$. The risk curves that are given in Fig. 3 are for $h = 1.7$ for which the sDOF was 226, but the minimum risk decision remained stable for a wide range of smoothing parameters. On the basis of the quadratic risk calculations QAIC chose six components and QBIC chose four components. Moreover using the MRA criterion it was clear that four components were adequate.

There is a sense in which selecting four components is the right answer for this problem. Biologists conjecture that if the interaction model is true there should be four groups of genes each displaying a separate profile under the four conditions. Among the three clusters that are given by the BIC, one profile matched with the distinct expression profile for $\Delta$-TAND with no mutation, but the other two gene clusters were mixed. In contrast the four clusters that were given by QBIC clearly identify four distinct patterns under the major combinations, thereby providing a meaningful summary and cluster analysis of these high dimensional data. Additionally QAIC and more prominently QBIC show a local minimum for a two-components mixture. This suggested that there might be two big groups. Further exploration shows that the two top clusters differentiated the mutated and unmutated TBP yeast strains.

6. Conclusion and future direction

In this paper we have developed a comprehensive tool for high dimensional model selection, using quadratic risk. Key to the calculation of quadratic risk is its representation and decomposition by using appropriate projection operators, and the estimation of these operators by using empirical versions. Decomposing the quadratic risk into the model lack of fit and the parameter estimation cost enabled us to build a method, QBIC, that mimics the BIC.

One feature of the derivation of our methodology, one that separates it from the AIC and BIC, is that the asymptotic expansions are based on non-parametric goodness-of-fit tests, not likelihood ratio asymptotics. We believe that our unified approach for building and estimating quadratic risk could pave the way for designing an appropriate model selection criterion for a host of previously unsolved problems, especially where the irregularity of parameter space eliminates the asymptotic theory underlying the use of the AIC, BIC and similar other criteria. In addition, the risk of the empirical estimate, which is a natural outcome of our risk analysis, provides a threshold that enables us to assess whether the model that was selected was itself a good fit. Unlike the AIC and BIC where only the optimum model is chosen, model adequacy provides us with the extra information of whether any of the models are adequate, or if there is a range of models that are adequate.

In addition, we also showed that the use of an appropriate kernel could enable us to minimize the computational burden that is associated with a high dimensional problem. Moreover, there is certainly more to learn about the structure of a data set than can be revealed by analysing data with a single model selection criterion. This is an area of future research.

Appendix A: Proof of proposition 1

We start with an asymptotic approximation to parameter estimation error $d_K(F_\tau, M_{\hat{\theta}}) - d_K(F_\tau, M_{\theta})$. Now we define a function in the $\theta$-space, by $L(\theta, \hat{\theta}) = d_K(M_{\hat{\theta}}, M_{\theta})$, which has the form

$$L(\theta, \hat{\theta}) = \int \int K(x, y) d(M_{\hat{\theta}} - M_{\theta})(x) dx d(M_{\hat{\theta}} - M_{\theta})(y).$$
For density function \( m_{\hat{\theta}} \) we assume that we can create a Taylor series expansion in \( \theta \) such that
\[
(m_{\hat{\theta}} - m_{\theta})(x) = (\hat{\theta} - \theta) \mathbf{u}_{\theta}(x) \mathbf{dM}_{\theta}(x) + o(|\hat{\theta} - \theta|).
\]
Thus we obtain the following quadratic approximation for the loss:
\[
m L(\theta, \hat{\theta}) = m(\hat{\theta} - \theta)^T \left\{ \int \mathbf{u}_{\theta}(x) K(x, y) \mathbf{u}_{\theta}(y) M_{\theta}(x) M_{\theta}(y) \right\} (\hat{\theta} - \theta) + o(m|\hat{\theta} - \theta|^2) \quad (36)
\]

Next, under standard regularity assumptions, it is a well-known result for maximum likelihood that
\[
m^{1/2}(\hat{\theta} - \theta) = m^{1/2} \frac{1}{m} \left\{ \sum_i \mathbf{u}_{\theta}(x_i) \right\} + o_p(1).
\]
This shows that the error term in equation (36) is \( o_p(1) \). In addition, substituting this expression into the first term gives the relationship
\[
m L(\theta, \hat{\theta}) = m \int \int V(x, y) \hat{F}(x) \hat{F}(y) + o_p(1) \quad (37)
\]
as needed for the proposition. Note that \( \int V(x, y) M_{\theta}(y) = 0 \), so the asymptotic mean for \( m L(\theta, \hat{\theta}) \) is
\[
E \left\{ \int \int V(x, y) \hat{F}(x) \hat{F}(y) \right\} = \text{trace}_{\mathbf{M}_{\theta}}(V).
\]

References