Smooth projected density estimation

Heather Battey  Han Liu
University of Bristol  Princeton University

May 11, 2014

Abstract

We introduce a new family of procedures, which we call smooth projection estimators, for multidimensional density estimation. These estimators are defined by a projection of a nonparametric pilot estimate onto a finite mixture class. Estimators from this family are attractive in that they offer both flexibility and the possibility of incorporating structural constraints in the pilot estimation stage, whilst possessing a succinct parametric representation. The latter property is of paramount importance when dealing with large datasets, which are now commonplace in many application areas. As a prototypical example, we consider in detail the case in which the output from a histogram pilot estimator is projected on the convex hull of spherical Gaussian densities. In a simulation study, we show that this particular smooth projection estimator outperforms many popular nonparametric estimators over a wide range of data generating scenarios. We also indicate the usefulness of the proposed methods through a classification task based on the estimated density functions.

1 Introduction

A great deal of recent research in statistics has been motivated by the opportunities and challenges spawned from the “data revolution”, a term that describes the proliferation of data emerging from dramatic improvements in data acquisition technologies. Whilst these vast reservoirs of information provide rich prospects for understanding in numerous fields, including molecular biology, neuroscience, space physics and economics, their benefits can only be realised if flexible tools are in place to distill the complex information contained in these colossal datasets.

The particular focus of this paper is density estimation in moderate and high dimensional settings. Let $Y_1, \ldots, Y_n$ be a random sample from a probability distribution that possesses Lebesgue density $f_0$ with unknown (possibly bounded) support on $\mathbb{R}^d$. When we are unwilling
to make assumptions on $f_0$, a popular estimator for $f_0$ is the kernel density estimator (KDE), which was initially proposed by Fix and Hodges (1951). Although the KDE works well in many low dimensional situations ($d \leq 3$), its performance deteriorates rapidly for a fixed sample size as the number of dimensions increases. Indeed, the KDE is just one example from the nonparametric family of density estimators, for which the well-documented result of Stone (1980) dictates that the minimax mean squared error convergence rate under standard smoothness assumptions is of order $O(n^{-4/(4+d)})$. Besides the dimension-dependent convergence rate, specification of the KDE involves a delicate choice of $d(d - 1)/2$ tuning parameters (bandwidths); whilst this task is critical to its performance, there is no universally accepted procedure for bandwidth selection in kernel density estimation.

The aforementioned features of the KDE and other fully agnostic density estimators motivate (and even necessitate) the use of procedures that are able to exploit hidden structure to make efficient statistical inference. Examples of these hidden structures include: sparsity, low-dimensional manifold, smoothness, copula structure, and conditional independence relations. The imposition of shape restrictions can also be an effective strategy for reducing the estimation error in high dimensions, even though this produces a degree of approximation error in the case that the restrictions are not satisfied. The last decade or so has witnessed heightened attention in estimators that are able to exploit such structural constraints. We provide in the next paragraphs a partial list of references relevant for the case of density estimation; further discussion is provided in §2.

In the case of graphical structure, nonparanormal graphical models (Liu et al., 2012a, 2009), transelliptic graphical models (Liu et al., 2012b), forest structures (Liu et al., 2011) and the density rodeo (Liu et al., 2007) have been proposed. Shape restrictions may also be imposed in a variety of forms, such as log-concavity (Cule et al., 2010), unimodality (Braun and Hall, 2001; Hall and Presnell, 1999) and $k$-monotonicity (Balabdaoui and Wellner, 2010). Inspection of any of the above references reveals that the use of such structural constraints in nonparametric estimation reduces the burden of choosing tuning parameters, or at the very least, weakens the sensitivity of such procedures to this delicate choice; as alluded to in our discussion of the KDE, without constraints, the selection of tuning parameters in nonparametric density estimation is a crucial yet monstrous task in high dimensions.

Until this point, we have focussed on nonparametric procedures and the potential improvements that are available from introducing structural constraints. However, it is important not to dismiss the computational burden and data storage requirements associated with nonparametric estimation on large datasets. Since nonparametric estimators are defined through an infinite dimensional parameter, estimates must be evaluated at a particular set of evaluation points, which are often either the data points themselves or a set of points associated with a dense and regular grid. This implies colossal data storage requirements when the data themselves are high dimensional. Moreover, in most practical situations, density estimation forms the basis of a decision making process, for which we require evaluation
of the density estimate at a new data point. This objective is clear if we consider the task of classifying a new patient into a disease group based on a high dimensional set of features. With this in mind, the ability to rapidly evaluate the density estimate associated with a new data point is a quality of paramount importance; in this respect, nonparametric estimators are impractical and the succinct low-dimensional representation of a parametric density is far more appealing.

We summarise the advantages and disadvantages of parametric and nonparametric density estimators in Table 1. It is natural to ask whether we can get the best of both worlds; on the one hand we would like the flexibility offered by a nonparametric estimator, but with the ability to incorporate constraints and exploit hidden structure. On the other hand, we would like faster convergence rates and the succinct representation of a parametric density estimator, which produces estimates that are easily stored and evaluated.

| Pros | Nonparametric (e.g. KDE) | Parametric (e.g. Finite mixture model) |
|------|--------------------------|---------------------------------------|
|      | • Agnostic               | • Ease of storage                      |
|      | • Easy to enforce structural constraint | • Rapid evaluation at new data point |
|      |                          | • Fast rate of convergence            |
| Cons | • Slow rate of convergence | • Hard to enforce structural constraints |
|      | • Difficulty of storage  | • Hard to justify a model             |
|      | • Hard to evaluate out-of-support data | |
|      | • Hard to choose tuning parameter | |

This paper proposes a new class of procedures for estimating a multidimensional density, whose aim is to achieve the best of both worlds: parametric and nonparametric. Our approach is useful for estimating densities in moderate dimensions (say, \( d \leq 10 \)), but it also addresses the fundamental challenge of estimating a multidimensional density when handling the huge datasets that routinely arise in many application areas.

2 Background

When the sample size is small relative to the dimension of the unknown object that we wish to estimate, traditional estimators are highly unstable, which necessitates regularisation. Regularisation can take various forms, but always involves structural assumptions on the object to be estimated. For instance, in the high dimensional regression model, sparsity assumptions are crucial for both stability and interpretability, whilst when the unknown object is a function (e.g. a density function), the term dimensionality plays two roles. On
the one hand, functions are infinite dimensional, in the sense that an infinite number of parameters are required to describe them; at the same time, the domain of the function can be high dimensional. Conventional nonparametric estimators, such as the KDE, are designed to deal with the first of these dimensionality concerns, and regularisation has traditionally been imposed through smoothness constraints on the function to be estimated. When the dimension of the domain is moderate or large, more regularisation may be necessary than that implied through a smoothness assumption, and this is achieved through structural constraints. We describe below what we mean by structural constraints.

2.1 Structure through graphical models

Analogous to the sparsity constraint in the high dimensional regression model, graphical models offer a powerful framework for exploring relationships between a large number of random variables. When random variables are multivariate Gaussian, a hidden sparsity structure can be uncovered through estimation of the covariance matrix, and there are many regularised estimators for a high dimensional covariance matrix that make use of such a sparsity assumption, these include thresholding (Bickel and Levina, 2008; El Karoui, 2008), penalty-based methods (Lam and Fan, 2009), and regularising principal components (Johnstone and Lu, 2009; Zou et al., 2006). Clearly however, if our objective is multidimensional density estimation, it is somewhat unnatural to impose a Gaussianity assumption in order to uncover a hidden sparsity structure.

Recently, graphical models have been proposed under the more flexible assumption a nonparanormal density (Liu et al., 2012a, 2009). A random vector \( Y = (Y_1, \ldots, Y_d)^T \) is said to possess a nonparanormal distribution (written \( Y \sim NPN(\mu, \Sigma, f) \)) if there exist functions \( \{f_j\}_{j=1}^d \) such that \( Z := f(Y) \sim N(\mu, \Sigma) \) where \( f(X) = (f_1(Y_1), \ldots, f_d(Y_d)) \). The scope of nonparanormal graphical models was further broadened by the introduction of transelliptical graphical models in Liu et al. (2012b). Both nonparanormal graphical models and transelliptical graphical models allow arbitrary graphical structures, but both involve an assumption on the class of distributions from which the data were drawn. An alternative approach is to restrict the graphs to being trees or forests, where each pair of vertices is connected by at most one path (see e.g. Liu et al., 2011); this places the constraints on the graph structure rather than on the class of distributions from which \( Y \) is drawn.

2.2 Structure through shape constraints

When the dimension \( d \) is moderate, or when the effective dimension uncovered through the use of a graphical model is moderate, shape constraints can be an effective form of regularisation. Shape-constrained estimation has become an active topic in statistics, yet, up until very recently, the case of \( d \geq 2 \) had received little attention. In the context of density
estimation, the log-concave maximum likelihood estimator (Cule et al., 2010) overcomes the theoretical and practical obstacles associated with such problems, and is an important and influential contribution. The imposition of the log-concavity constraint allows the authors to completely circumvent the difficult problem of selecting tuning parameters, which is a crucial feature of all nonparametric procedures that do not make use of shape constraints, and is thus particularly appealing in multidimensional scenarios. A noteworthy point is that the omission of tuning parameters results in estimates that are non-differentiable.

Beyond log-concavity, a variety of shapes can be imposed in the KDE through the tilting method of Hall and Presnell (1999) (see also the discussion of Braun and Hall (2001)). Again, the imposition of shape constraints weakens the sensitivity of the estimator to the delicate choice of tuning parameters.

2.3 Difficulty incorporating constraints in mixture classes

Density estimation in high dimensions is often performed through mixture models. Mixture models with a fixed number of components are parametric, and thus are attractive in that they provide a succinct representation for the estimated density, which may be stored easily and evaluated rapidly. However, choosing the number of mixtures is the same as choosing a parametric model, and the flexibility offered by a large mixture class inevitably results in overfitting. Although nonparametric mixture models (see e.g. Richardson and Green, 1997), which are able to adapt to the number of mixtures, are able to address this difficulty, it is not at all clear how structural constraints can be imposed in parametric or nonparametric mixture modelling.

2.4 Structure and succinctness through projection

The key idea of this paper is to make use of the succinct representation of a finite mixture class (i.e. a parametric class), whilst preventing overfitting and potentially exploiting structural constraints. We do this by projecting a nonparametric pilot estimator of the practitioner’s choice on a finite mixture class (also of the practitioner’s choice); we formalise this idea in §3.2 and provide a concrete example in §3.3. Although the nonparametric pilot estimator may converge slowly, the projection step is a form of smoothing, and thus intuitively can lead to faster rates.
3 Methodology

3.1 Notation

Let $F_S$ be an arbitrary mixture class indexed by a Euclidean parameter space $\Theta$, let $g$ denote an arbitrary density function (not necessarily Lebesgue measurable) on $\mathbb{R}^d$, and let $Q$ denote a probability measure on $\mathbb{R}^d$. With notation inspired by Samworth and Yuan (2012), the projection operator is defined by

$$
\psi_{F_S}^*(Q, g) := \arg\min_{f \in F_S} \int_{\mathbb{R}^d} (g - f)^2 dQ.
$$

(3.1)

Introduce $f_0$, $f_n$ and $\hat{f}_P$, which denote respectively, the true density function, the collection of $(1/n)$-weighted point masses at the points where the data are observed, and a generic nonparametric pilot density estimator. We will later refer to a concrete example in which $\hat{f}_P$ is taken as $\hat{f}_H$, a histogram pilot density estimator. We will also refer extensively to the case in which $F_S = \bar{F}_S$, where $\bar{F}_S$ is a spherical Gaussian mixture class defined as

$$
\bar{F}_S := \left\{ f \in F_d : f(y) = \sum_{s=1}^S \pi_s \phi(y; \mu_s, q I_d) : \pi \in \Delta^S, \mu_s \in \mathbb{R}^d \forall s \in \{1, \ldots, S\} \right\},
$$

(3.2)

with $\Delta^S$ the $S$-dimensional unit simplex. Two other function classes of interest in this paper are the infinite spherical Gaussian mixture class,

$$
\bar{F}_d := \left\{ f \in F_d : f(y) = \int_{\mathcal{M}} \phi(y; \mu, q I_d) dG(\mu) : G \in \mathcal{G} \right\}
$$

(3.3)

where $\mathcal{M} = [-M, M]^d$ for $M < \infty$ and $\mathcal{G}$ is the class of all probability measures on $\mathcal{M}$, as well as the class of log-concave densities,

$$
F_{d}^{LC} := \{ f \in F_d : h = \log(f) \text{ is a concave function} \}.
$$

(3.4)

In light of (3.1), the following projections are of interest:

$$
f_0^* := \psi_{\bar{F}_d}^*(P, f_0) \quad \hat{f}_P^* := \psi_{\bar{F}_d}^*(P, \hat{f}_P) \quad f_n^* := \psi_{\bar{F}_d}^*(P, f_n) = \psi_{\bar{F}_d}^*(\bar{P}_n, \hat{f}_P)
$$

as well as that defined through

$$
\hat{f}_n^* := \psi_{\bar{F}_d}^*(\bar{P}_n, \hat{f}_P)
$$

where $P$ is the true distribution from which the data were drawn and $P_n$ is the empirical distribution.

Although it will usually be convenient to make use of the compact notation of projection operators, it will occasionally be necessary to refer to the $L_2(P_n)$-criterion function that corresponds to $\psi_{\bar{F}_d}^*(P_n, g)$,

$$
L_n(f; g) = \int_{\mathbb{R}^d} (g - f)^2 dP_n.
$$

(3.5)
3.2 General framework

Given a sample $Y_1, \ldots, Y_n$ of $n$ observations drawn randomly from the true density $f_0$ in $\mathbb{R}^d$, we construct an estimate of the density as the corresponding realisation of the pilot estimator $\hat{f}^P$. The choice of pilot estimator may be one of convenience; more interestingly, it may exploit some assumed structural properties of $f_0$, such as those discussed in §2, or be able to detect and exploit hidden structure. The resulting estimate is characterised by the advantages and disadvantages of the left hand column of Table 1; in particular, it may be non-differentiable, and awkward to evaluate and store.

To remove these undesirable features, we then project our pilot estimate on the convex hull of a large dictionary of parametric densities parameterised by $\Theta \subseteq \mathbb{R}^k$ for some $k \geq 1$, i.e. the mixture class $\mathcal{F}_d^S$. This mixture class is indexed by $\Theta = \{(\xi, \pi) : \pi \in \Delta^S, \xi \in \Xi\}$, hence the smooth projection estimator,

$$\hat{f}_n^* := \psi^*_{\mathcal{F}_d^S}(P_n, \hat{f}^P),$$

may be equivalently written as $\hat{f}_n^* := f_n^*(\hat{\xi}, \hat{\pi})$, where $(\hat{\xi}, \hat{\pi})$ are solutions to the finite dimensional minimisation problem

$$\left(\hat{\pi}, \hat{\xi}\right) := \arg\min_{\pi \in \Delta^S, \xi \in \Xi} \sum_{i=1}^n \left(\hat{f}^P_n(Y_i) - f(Y_i; \pi, \xi)\right)^2, \quad f \in \mathcal{F}_d^S. \quad (3.7)$$

Although equation (3.6) is the most natural way to define our projection estimator, another possibility, which may be attractive in small $n$ scenarios, is to replace $P_n$ by $P^B_J$ where $J > n$ and $P^B_J$ is the empirical measure of the points $U_1, \ldots, U_J$, each uniformly distributed over $[-B, B]^d$, $B < \infty$.

3.3 Concrete example

3.3.1 Basic construction and properties of the histogram

Suppose that $Y_1, \ldots, Y_n$ are independent copies of a random variable $Y$ whose distribution possesses some Lebesgue density $f_0$. We specify a partition, $\mathcal{P}_n$, of $\mathbb{R}^d$ through an anchor point (taken as the origin without loss of generality) and a collection of bin widths $\{h_{n,j} : j = 1, \ldots, d\}$. The (non-adaptive) histogram density estimator is defined as

$$\hat{f}^H(y) = \frac{1}{n \prod_{j=1}^d h_{n,j}} \sum_{i=1}^n \mathbb{1}\{Y_i \in A_n(y)\}$$

(3.8)

where $A_n(y)$ is the set in $\mathcal{P}_n$ that contains $y$. It is well known (see e.g. Devroye et al., 1996, and references therein) that $\hat{f}^H$ converges to $f_0$ almost surely in $L_1$ norm provided that the bin widths $(h_{n,j})_{j=1}^d$ are chosen in such a way that $\max_j h_{n,j} \to 0$ and $n \prod_{j=1}^d h_{n,j} \to \infty$ as
More refined theory for optimal $L_1$ bin width selection has been established in Devroye and Lugosi (2004), whilst Kanazawa (1992) discusses Hellinger optimality criteria and Stone (1985) established $L_2$ near-optimality of cross validation over the class of bounded densities. More advanced (adaptive) bin width selection procedures have also been widely studied. These procedures allow the bin widths to become wider in the tails, thereby allowing for densities whose tails decay slowly.

Despite the many desirable features and the appealing asymptotic theory that accompanies the histogram density estimator, it remains a rather unpopular choice in practice due to the piecewise constant (thus non-differentiable) nature of the resulting density estimates. This is not necessarily a problem per se, but it can have negative consequences when the density estimator is used as a preliminary step in a more sophisticated procedure, for instance when the density estimates are used to define a Bayes’ classifier. A further drawback of the histogram density estimator is its property of being less accurate at the borders of the cells of the partition than at the cell centres. This was the motivation behind the kernel density estimator, which was based on the observation that points near the border of a cell should be less influential in a decision regarding the cell’s centre. Our smoothing strategy exploits this observation by using the points at which the data are observed to define the least squares projection of the histogram on a dictionary of parametric densities. It is easy to see heuristically, that the probability of observing data points at the edges of the histogram cells is lower than the probability of observing data that fall within the cells, thus the histogram bin centres play a more influential role in the criterion function than do the bin edges.

3.3.2 Projection onto the class of location mixtures of spherical Gaussians

Just as the choice of pilot density estimator is somewhat arbitrary within the confines of the nonparametric class, so too is the choice of parametric family onto which we project, and this may be dictated by prior knowledge or by computational or algebraic convenience. In our simulation experiments of §5.1, we choose to project the histogram pilot onto the class of location mixtures of spherical Gaussians, defined in equation (3.2) above. Hence the minimisation problem in (3.7) may be re-written as

$$\left( (\hat{\pi}_s)_{s=1}^S, (\hat{\mu}_s)_{s=1}^S \right) = \arg\inf_{\pi \in \Delta^S, (\mu_s)_{s=1}^S \in \mathbb{R}^d} \sum_{i=1}^n \left( \hat{f}^H(Y_i) - \sum_{s=1}^S \pi_s \phi(Y_i; \mu_s, \frac{q}{2}I_d) \right)^2.$$  

(3.9)

where the choice of $S$ and $q$ is further discussed in §3.3.3 and §6.

3.3.3 Computation of the mixture parameters and weights

In the usual unsupervised framework, estimation of the mixing parameters $\pi_1, \ldots, \pi_S$ and the mean vectors $\mu_1, \ldots, \mu_S$ of a Gaussian mixture model is usually performed using the EM
algorithm (Dempster et al., 1977), which alternates between performing an expectation (E) step and a maximisation (M) step. It seeks to compute the maximum likelihood estimates by maximising the expected likelihood obtained in the E step. The parameters found in the M step are then used to perform another E step, and the process is repeated until convergence. Although Dempster et al. (1977) proved that the algorithm does converge, there is no guarantee that it converges to the maximum likelihood solution. Indeed, when the scale parameters are not fixed in advance, the likelihood is unbounded and the maximum likelihood solution does not exist.

In a similar spirit, we propose that estimation of $\pi_1, \ldots, \pi_S$ and $\mu_1, \ldots, \mu_S$ be performed in our supervised regression framework by a least-squares alternation algorithm, similar to that proposed by Yuan (2009). The algorithm iteratively minimises equation (3.9) with respect to the mixing proportions, with the $\{\mu_s : s = 1, \ldots, S\}$ held fixed (at their estimated values, or at some initial starting values in the first iteration), and then with respect to the $\{\mu_s : s = 1, \ldots, S\}$ with the mixing proportions held at their estimated values at the previous iteration. These steps are repeated until the difference in the objective function from the two minimisation steps is within some small tolerance level. With one set of unknowns held fixed, the minimisation can be performed efficiently with standard quadratic program solvers. We suggest taking initial $\pi_1, \ldots, \pi_S$ as the center point of the unit $S$-simplex, i.e. $(1/S, \ldots, 1/S)$. The quadratic program solver for $\mu_1, \ldots, \mu_S$ also requires a set of starting values for $\mu_1, \ldots, \mu_S$ in the first iteration; close inspection of the proof of Lemma 7.2 in §7 reveals that a judicious choice of initial $\mu_1, \ldots, \mu_S$ is to take them equally spaced in $\mathcal{M} = [-M, M]^d$. We may use this to guide our choice of $S$, as taking an $S$ that possesses an integer-valued $d^{th}$ root allows us to place the $\mu_1, \ldots, \mu_S$ on a regular grid over a pre-defined boundary, which, ideally, is close to $[-M, M]^d$ itself. We suggest using an unconstrained solver such as MATLAB’s fminunc for minimisation with respect to $\mu_1, \ldots, \mu_S$ and a constrained solver such as MATLAB’s fmincon for minimisation with respect to $\pi_1, \ldots, \pi_S$. Although, in theory, we could specify bounds of $[-M, M]$ on each of the $\mu_1, \ldots, \mu_S$ vectors, we prefer not to do this since $M$ is unknown in practice.

4 Theoretical results

Lemma 4.1. Let $(U_1, \ldots, U_J)$ be independent random vectors in $[-B, B]^d$ for some $B < \infty$, let $\mathcal{G}$ be the class of all probability measures on $[-M, M]^d$ and let $g$ be an arbitrary density function (not necessarily Lebesgue measurable) on $\mathbb{R}^d$, where $M$ is that of (3.3). With probability 1, the minimum of

$$Q_J(G) := \frac{1}{J} \sum_{j=1}^{J} (g(U_j) - f(U_j; G))^2 \quad f \in \bar{F}_d, \quad G \in \mathcal{G}$$

(4.1)
exists, and there exists a distribution \( \hat{G} \) whose support contains no more than \( J + 1 \) points that achieves this minimum.

**Proof.** The existence of a minimum is guaranteed by the convexity of the objective function and the feasible region. Recall that

\[
F_d := \left\{ f \in F_d : f(y) = \int_M \phi(y; \mu, qI_d) dG(\mu) : G \in \mathcal{G} \right\}.
\]

Introduce

\[
A := \{ \phi(\cdot; \mu, qI_d) : \mu \in \mathbb{R}^d \}
\]

and observe that \( F_d \subseteq \text{conv}(A) \) where \( \text{conv}(\cdot) \) denotes the convex hull. Letting \( \hat{G} \) be a minimiser of (4.1), \( f(\cdot; \hat{G}) \) is an element of \( F_d \) and therefore also of \( \text{conv}(A) \).

Since \( \{f(U_j; \hat{G}) : j = 1, \ldots, J\} \in \text{conv}(A) \), we invoke Carathéodory’s convex hull theorem (see e.g. Rockafellar, 1997, Theorem 17.1) and conclude that there exists a subset \( \mathcal{B} \) of \( \mathcal{A} \) consisting of \( S \leq J + 1 \) points such that \( \{f(U_j; \hat{G}) : j = 1, \ldots, J\} \in \text{conv}(\mathcal{B}) \). i.e. there exists \( \mu_1, \ldots, \mu_S \in \mathbb{R}^d \) and \( \pi_1, \ldots, \pi_S \in \Delta^S \) such that

\[
f(U_j; \hat{G}) = \sum_{s=1}^{S} \pi_s \phi(U_j; \mu_s, qI_d) \quad \forall j \in \{1, \ldots, J\}
\]

where \( \Delta^S \) is the \( S \)-dimensional unit simplex. \( \square \)

**Corollary 4.2.** Let \( f_0 \) be a density supported on a compact subset of \( \mathbb{R}^d \) and let \( Y_1, \ldots, Y_n \) be independent draws from \( f_0 \). Then, with probability 1, the minimum of

\[
Q_n(G) := \frac{1}{n} \sum_{i=1}^{n} (g(Y_i) - f(Y_i; G))^2 \quad f \in \bar{F}_d, \quad G \in \mathcal{G}
\]

exists, and there exists a distribution \( \hat{G} \) whose support contains no more than \( n + 1 \) points that achieves this minimum.

**Proof.** The proof is a straightforward application of Lemma 4.1. \( \square \)

**Remark 4.3.** Corollary 4.2 can be relaxed to allow non-compact support by introducing a tail condition on the true law of \( Y_i \). Should we not wish to impose neither a tail condition nor a compact support condition, Corollary 4.2 is still applicable (mutatis mutandis) if \( g \) is taken as a nonparametric pilot estimator \( \hat{f}^P \) whose performance does not rely on tail or support conditions and if we sample randomly from \( \hat{f}^P \) over an arbitrarily large compact subset of \( \mathbb{R}^d \).
Theorem 4.4, whose proof appears in §8, formalises the statement that projection of the pilot estimate on the mixture class allows faster rates of convergence than those achievable by the pilot estimator itself. Note however, that convergence of the smooth projection estimator is to the projection of the true density function on the mixture class rather than to the true density function itself.

Theorem 4.4. Suppose that there exists a \( \theta^*_P = (\pi^*_P, \xi^*_P) \) such that \( \hat{f}^*_P := f(\theta^*_P) \) is the unique minimiser of \( \int (\hat{f}^P - f)^2 dP \) over \( \mathcal{F}^*_d \) and suppose \( \hat{f}^P \) satisfies \( r_n(\hat{f}^P - f_0) \rightsquigarrow X \) for some separable \( X \). Then

\[
\left\| \psi^*(P_n, \hat{f}^P) - \psi^*(P, f_0) \right\|_2 = o_p \left( r_n^{-1} \right).
\]

5 Numerical results

This section makes use of the estimator defined by our concrete example in §3.3. In the first subsection, we consider artificial data generated from three example densities. The first two densities are standard whilst the third is chosen to represent a particularly challenging density estimation problem; one in which the true density has bounded support over a non-convex set. In \( \mathbb{R}^2 \) and \( \mathbb{R}^3 \), ecological statistics abounds with such data; see for example the Aral sea chlorophyll population study considered in Wood et al. (2008), or the Blue Ling population study of Augustin et al. (2013). Our prototypical example considers data that arise as realisations of a random variable that is normally distributed around the circumference of a circle.

5.1 Simulation performance

We obtain 1000 pseudo random samples of size \( n = 50, 100, 250, 500 \) from a bivariate density \( f_0 \). We consider the following densities for \( f_0 \): (a) an equally-weighted mixture of two scale location normal densities with mean vectors \( \mu_1 = (1,2)^T, \mu_2 = (-1,1)^T \) and covariance matrices \( \Sigma_1 = [2, -0.5; -0.5, 1.5], \Sigma_2 = [4, 0.9; 0.9, 1.5] \); (b) a bivariate gamma density with the same independent marginals (with shape parameter equal to 2 and scale parameter equal to 1); (c) an equally weighted mixture of 200 uncorrelated bivariate normal densities with scale 0.7 and location on a radius 4 ring in \( \mathbb{R}^2 \). (a) is a skewed unimodal density exhibiting dependence, (b) is a unimodal density that readily admits structural constraints, and (c) is a density supported on a compact, but non-convex domain. Figure 1 illustrates three different versions of the histogram constructed in a typical realisation of \( n = 100 \) draws from the density in case (b). The first histogram is the perturbed unconstrained one, obtained as a mixture of 5 histograms of fixed bin width each centered at a different anchor point. The second histogram is obtained through a random forest procedure, which is equivalent to estimating the histogram for each independent marginal and multiplying them together in
order to form the bivariate density. Finally, the third histogram is constructed as the second histogram, but additionally makes use of a unimodality constraint. For the independent marginals of each component histogram in the mixture, the unimodal histogram moves outwards from the modal bin and tests whether more extreme bins contain fewer observations than bins closer to the mode. In the case that they do not, the unimodal histogram moves the bin edges outwards until the condition is satisfied; this illustrates very clearly the well known result that shape constraints weaken the dependence on tuning parameters. As with the random forest histogram, the independent marginals are then multiplied together in order to form the bivariate density estimate.

Figure 1: Panel (A) is the true bivariate density with the same independent \( \Gamma(2,1) \) marginals, whilst panels (B)-(C) are versions of the histogram based on a sample of size \( n = 100 \) and a binwidth of \( 1.6(IQ_j)h_n \) in each coordinate direction, where \( (IQ_j) \) is the interquartile range in the \( j^{th} \) coordinate direction and \( h_n = n^{-1/2d} \) with \( d = 2 \). Panel (B) is the perturbed unconstrained histogram (mixture of 5 histograms around a perturbed anchor point); panel (C) is the random forest histogram; panel (D) is the perturbed histogram with a unimodality constraint on each of the histograms in the mixture.

Figure 2 illustrates the estimates constructed by several different estimators based on a typical draw from the density of case (c). To be more explicit about the data generation for case (c), let \( t \) be a vector of 500 equally spaced points in \([0, 2\pi]\). We generate artificial data normally distributed around the circumference of the circle of radius 4, i.e. \( \{Y_i = (Y_{i,1}, Y_{i,2}) : i = 1, \ldots, 500\} \) where \( Y_{i,1} = 4 \cos(t_i) + \epsilon_i \) and \( Y_{i,2} = 4 \sin(t_i) + \epsilon_i \) where \( \epsilon_i \sim N(0, 0.7) \).
Figure 2: Panel (A) is the true density constructed as an equally weighted mixture of 200 uncorrelated bivariate normal densities with scale 0.7 and location on a radius 4 ring in $\mathbb{R}^2$. Panels (B)-(F) are estimated based on the same $n = 250$ observations, more specifically: panel (B) is the log concave maximum likelihood estimate of Cule et al. (2010); panel (C) is the least-squares cross-validated bandwidth kernel density estimate, computed in $\mathbb{R}$ using the ks package (Duong, 2007); panel (D) is the kernel density estimate using a 2-stage plug-in bandwidth estimate (see Duong and Hazelton, 2003; Wand and Jones, 1995; Sheather and Jones, 1991); panel (E) is an equally weighted mixture of 5 perturbed histograms each constructed using a bin width of $(IQ_j)h_n$ in each coordinate direction, where $(IQ_j)$ is the interquartile range in the $j^{th}$ coordinate direction and $h_n = n^{-1/2d}$ with $d = 2$; panel (F) is the projection on the class of location mixtures of spherical Gaussians with $\varphi = 0.7$ and $S = 64$.

Figure 3 illustrates the behavior of the integrated squared error (ISE) with increasing
sample size for the density of scenario (a). We consider (i) the kernel density estimator with
least-squares cross-validated bandwidth (Wand and Jones, 1995; Duong, 2007); (ii) the kernel
density estimator with 2-stage plug-in bandwidth (Duong and Hazelton, 2003; Wand and
Jones, 1995; Sheather and Jones, 1991); (iii) the log concave maximum likelihood estimator of
Cule et al. (2010); (iv) the unperturbed histogram pilot with bandwidth $1.6(IQ_j)h_n$ in each
coordinate direction, where $(IQ_j)$ is the interquartile range in the $j^{th}$ coordinate direction
and $h_n = n^{-1/2d}$ with $d = 2$; (iv) the projection of (iii) on the finite location mixture
of spherical Gaussians, with $q = 1$ and $S = 36$; (v) the mixture of five histograms, each
constructed as in (iii) but with anchor points that are perturbations of the zero vector in
different coordinate directions; (vi) the projection of (v) on the finite location mixture of
spherical Gaussians, with $q = 1$ and $S = 36$.

Figure 3: Location-scale normal mixture example. Left segment to right segment correspond to
$n \in \{50, 100, 250, 500\}$. ‘K-CV’ is the cross validated kernel density estimator; ‘K-PI’ is the plug-in
bandwidth kernel density estimator; ‘Lcd’ is the log-concave maximum likelihood estimator; ‘Hist’
is the histogram estimator based on a single anchor point at the origin; ‘Project’ is the $L_2(P_n)$-
projection of ‘Hist’; ‘pHist’ is the mixture of 5 histograms each based on an anchor point that is a
perturbation of the origin; ‘pProject’ is the $L_2(P_n)$-projection of ‘pHist’.

Figure 4 is the analogue of Figure 3 for scenario (b), except that we consider some
additional estimators: (vii) the random forest histogram that exploits the independence
of the marginals; (viii) the projection of (vii) on the finite location mixture of spherical
Gaussians, with $q = 1$ and $S = 36$; (ix) the histogram that exploits the independence of
the marginals and the unimodality in the directions of the standard axis in $\mathbb{R}^2$ (x) the projection
of (ix) on the finite location mixture of spherical Gaussians, with \( q = 1 \) and \( S = 36 \). Figure 5 is the analogue of Figure 3 for scenario (c), except that we increase the number of mixture components in the projection to \( S = 64 \) in order to account for the difficulty in estimating a density that is supported on a ring, we also drop the histogram pilot bandwidth to \((IQ_j)h_n\) and take \( q = 0.7 \) in the mixture class.

**Figure 4**: Gamma example. Left segment to right segment correspond to \( n \in \{50, 100, 250, 500\} \). ‘K-CV’ is the cross validated kernel density estimator; ‘K-PI’ is the plug-in bandwidth kernel density estimator; ‘Lcd’ is the log-concave maximum likelihood estimator; ‘Hist’ is the histogram estimator based on a single anchor point at the origin; ‘Project’ is the \( L_2(P_n) \)-projection of ‘Hist’; ‘pHist’ is the mixture of 5 histograms each based on an anchor point that is a perturbation of the origin; ‘pProject’ is the \( L_2(P_n) \)-projection of ‘pHist’; ‘forestHist’ is the mixture of 5 histograms that additionally exploit the information that the marginals are independent; ‘forestProject’ is the \( L_2(P_n) \)-projection of ‘forestHist’; ‘uForestHist’ is the mixture of 5 histograms that exploit the information that the marginals are independent and unimodal; ‘uForestProject’ is the \( L_2(P_n) \)-projection of ‘uForestHist’.
Figure 5: Ring example. Left segment to right segment correspond to \( n \in \{50, 100, 250, 500\} \). ‘K-CV’ is the cross validated kernel density estimator; ‘K-PI’ is the plug-in bandwidth kernel density estimator; ‘Lcd’ is the log-concave maximum likelihood estimator; ‘Hist’ is the histogram estimator based on a single anchor point at the origin; ‘Project’ is the \( L_2(P_n) \)-projection of ‘Hist’; ‘pHist’ is the mixture of 5 histograms each based on an anchor point that is a perturbation of the origin; ‘pProject’ is the \( L_2(P_n) \)-projection of ‘pHist’.

Figures 3, 4 and 5 provide numerical evidence that our smooth projection procedure is universally unintimidated across a range of data generating mechanisms. For the true density of scenario (a), the smooth projection estimator dramatically outperforms the other estimators for all sample sizes. For the bivariate gamma density of scenario (b), the smooth projection estimator based on the perturbed histogram achieves a smaller median integrated squared error than its competitors for small sample sizes, and remains competitive for larger sample sizes, whilst being eventually dominated in terms of median integrated squared error by the log-concave maximum likelihood estimator, whose log-concavity constraint is appropriate for case (b). A similar situation emerges for the ring example of scenario (c), but here the log-concave maximum likelihood estimator is not competitive, even in large samples, due to the imposition of the log-concavity constraint. This constraint leads to a density estimate that is supported on the convex hull of the data, and thus cannot capture the lack of mass in the centre of the ring; this is illustrated in panel (B) of Figure 2. Figure 4 indicates that the use of structural constraints can be advantageous in small and moderate sample sizes, however, due to the relatively restrictive nature of the mixture class onto which we project, the improvement in estimation error from inclusion of the constraint in the histogram pilot.
estimates is not directly translated to the projection.

5.1.1 Comparison to EM in the presence of structural information

To study the value of incorporating structural constraints in the pilot estimation stage, we compare the smooth projection estimator to the Kiefer-Wolfowitz estimator when the true density possesses structure that we may detect and exploit in the pilot estimation stage. More specifically, for each of $S \in \{9, 16, 25, 36, 49\}$ we draw (using the same seeds) 1000 samples of size $n = 250$ from a bivariate gamma distribution with independent $\Gamma(4, 0.75)$ marginals. We use the ‘forestProject’ version (cf Figure 4 and surrounding text) of the smooth projection estimator (labelled ‘pFP’ in Figure 6) and for the Kiefer-Wolfowitz estimator we fix the same spherical correlation structure and the same scale $q = 1$ and estimate the mixture density using the EM algorithm (labelled ‘EM’).

![Figure 6: Performance of the smooth projection estimator and EM implementation of the Kiefer-Wolfowitz estimator for $f_0$ the density of a bivariate gamma distribution with independent $\Gamma(4, 0.75)$ marginals. Panels correspond (from left to right) to 9 mixtures, 16 mixtures, 25 mixtures, 36 mixtures, 49 mixtures.]

5.1.2 Robustness to bin width choice in the pilot estimation

As alluded to in §1, the choice of tuning parameters in nonparametric density estimation is of crucial importance to the performance of the estimator. This point is emphasised in Figures 3, 4 and 5, which highlight the dramatic performance discrepancies between the KDE with
cross validated bandwidth and the KDE with plug-in bandwidth. This observation induces us to ask whether the projection step can robustify density estimates to the delicate choice of tuning parameters in the nonparametric pilot estimation stage. Figure 7, which is based on 250 draws from the density in scenario (a) (using the same seeds to generate 100 Monte Carlo replications for each bin width parameter), answers this question in the affirmative. Whilst the performance of the histogram and perturbed histogram are both affected substantially by the choice of bin width, the performance of the estimate given by the projections onto the class of location mixtures is relatively stable across the different bin width choices. This is an important feature of our procedure, as the choice of tuning parameter can be an extremely challenging problem, especially in high dimensions.

![Figure 7: Robustness to changes in the histogram bin width. From left segment to right segment: increasing constant in the histogram bin width from 0.6 to 2.2 in the location-scale mixture of normals example for a fixed sample size of $n = 250$ and a fixed $q$ of 0.7.](image)

### 5.2 Real data example

A dataset that is commonly considered in the context of multivariate density estimation (see e.g. Liebscher, 2005; Cule et al., 2010) is the Wisconsin breast cancer (diagnostic) dataset, publicly available on the UCI Machine Learning Repository website: [http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29](http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29).

It consists of 30 real-valued continuous attributes based on the cell nuclei of 569 breast tumor patients, of which 212 instances are malignant and 357 instances are benign, along with a variable indicating whether the tumor was malignant or benign. The dataset is discussed in more detail in Street et al. (1993). As in Cule et al. (2010), we consider the first two principal components of the 30 attributes, which accounts for 63% of the variability in the dataset, and we address the problem of trying to diagnose the individuals based on observations on the
scores associated with these first two principal components. Figure 7 provides a scatterplot of the data, together with the perturbed histogram estimate and the smooth projection estimate for the benign and malignant groups individually.

![Figure 7: Scatterplot of data with perturbed histogram and smooth projection estimates.](image)

**Figure 5:** From top left to bottom right: the first two principal component scores for the malignant and benign cases; cross validated kernel density estimates; average of 5 histograms around a perturbed anchor point with binwidth constant $c = 2$; projection of the histogram estimate on $F_d^S$ with $S = 36$ and $q = 1$.

We construct a test set of 50 randomly drawn observations from the sample of 569 observations; amongst these test observations, $n_M^o$ cases correspond to malignant tumours and $n_B^o$ to benign. On the remaining 519 observations, we construct bivariate density estimates based on training sets of $n_M$ patients in the malignant group and $n_B$ patients in the benign group using the version of our density estimator discussed in §3.3 as well as three competing density estimators: the KDE with cross validated bandwidth, the KDE with plug-in bandwidth, and the log-concave maximum likelihood estimator. Notice that $n_M^o, n_B^o, n_M$ and $n_B$ are random variables.

Let $\hat{f}$ denote an arbitrary density estimator and define the posterior probabilities associated with a Bayes' classifier by

$$
\hat{P}(1|y) = \frac{\hat{f}^1(y|1)\hat{P}(1)}{\hat{f}(y|1)\hat{P}(1) + \hat{f}(y|0)\hat{P}(0)} \quad \text{and} \quad \hat{P}(0|y) = \frac{\hat{f}^0(y|0)\hat{P}(0)}{\hat{f}(y|1)\hat{P}(1) + \hat{f}(y|0)\hat{P}(0)}.
$$

(5.1)

where $\hat{f}(y|1)$ and $\hat{f}(y|0)$ are density estimates obtained from the training observations in the malignant and benign group respectively, and $\hat{P}(1)$ and $\hat{P}(0)$ denote the estimated
probability of being in the malignant and benign groups respectively; these quantities are obtained using the empirical proportions of malignant and benign cases in the training set (i.e. \( n_M/(n_M + n_B) \) and \( n_B/(n_M + n_B) \) respectively). Letting \( \{ y_i : i = 1, \ldots, 50 \} \) denote the vector of observations for patient \( i \) in the test set, the relative magnitudes of \( \hat{P}(1|y_i) \) and \( \hat{P}(0|y_i) \) determine whether individual \( i \) with attributes associated with principal component score vector \( y_i \) is assigned to group 1 (malignant) or group 0 (benign). More specifically, individual \( i \) is classified into group \( j \in \{0,1\} \) if \( \hat{P}(j|y_i) > \hat{P}(\{j\}^{c}|y_i) \). If the estimated posterior probabilities of a patient being the malignant and benign groups are equal, we use the pessimistic default rule of classifying the patient into the malignant group.

We repeat the above experiment 5 times, recording the misclassification rates for four competing estimators in each case, i.e. letting \( m^o \) denote the total number of misclassifications and \( m^o_M \) denote the number of misclassifications in the malignant group, the misclassification rates recorded are \( m^o/50 \) and \( m^o_M/n^o_M \). We average these misclassification rates over the 5 experiments. The following table summarises the results for the KDE with cross validated bandwidth, the KDE with plug-in bandwidth, the log-concave maximum likelihood estimator, and our smooth projection estimator using the projection of the perturbed histogram pilot with bandwidth \( 2(IQ_j)h_n \) in each coordinate direction, where \( (IQ_j) \) is the interquartile range in the \( j \)th coordinate direction and \( h_n = n^{-1/4} \), \( q = 1 \) and \( S = 36 \).

|                  | KDE(CV) | KDE(PI) | LogConcave | SmoothProject |
|------------------|---------|---------|------------|---------------|
| Average \( m^o/50 \) | 0.0800  | 0.0840  | 0.0680     | 0.1000        |
| Average \( m^o_M/n^o_M \) | 0.1110  | 0.1201  | 0.0959     | 0.0921        |

Table 2: Out-of-sample misclassifications made by the Bayes’ classifier based on the indicated density estimator.

In this small experiment, the log concave maximum likelihood estimator performs best according to the criterion of minimising the total number of out-of-sample misclassifications. In terms of misclassifying patients whose tumors are of malignant type (arguably the more serious error), the Bayes’ classifier based on our smooth projection estimator performs best, slightly outperforming the log concave maximum likelihood estimator.

6 Discussion

6.1 Theoretical and empirical justification for the pilot estimation stage

Although, as argued in §1, the pilot estimation stage is useful in that it allows structure to be exploited, an important feature of the smooth projection estimator is that the projection stage will not in general preserve the structural constraints exploited in the pilot estimation.
A natural question that emerges from this observation is whether a better estimator can be constructed by projecting \( f_n \) directly on a mixture class that satisfies the structural constraints; we denote this class \( \mathcal{F}_{S,C} \). There are several reasons why this is not sensible. From a practical perspective, the implementation requires specification of \( \mathcal{F}_{S,C} = \mathcal{F}_d^S \cap \mathcal{F}_d^C \) for the particular constrained class of interest \( \mathcal{F}_{d}^C \); this may be difficult in many cases. Moreover, the additional constraint will lead to further constraint-specific algorithmic complications. The theoretical grounds for not seeking to do this are even more pervasive. Elements of \( \mathcal{F}_d^S \) are simply used to provide a succinct approximation to the density, if \( f_0 \notin \mathcal{F}_d^S \) then a fortiori \( f_0 \notin \mathcal{F}_d^S \cap \mathcal{F}_d^C \) even when the constraints are satisfied. The following example clearly illustrates the potential dangers of such an approach.

**Example 6.1.** Suppose \( f_0 \in \mathcal{F}_{d}^{LC} \) but \( f_0 \notin \mathcal{F}_d^S \). A proof of the following result is provided in §8.

**Proposition 6.2.** For some finite positive constant \( q \).

\[
\mathcal{F}_{d}^{LC} \cap \mathcal{F}_d^S = \left\{ f \in \mathcal{F}_d^S : \sum_{s=1}^{S} \phi(y; \mu_s, qI) \| \rho_s \| \leq 1 \quad \forall \ y \in supp(f) \right\},
\]

where \( \rho_s := (y - \mu_s)/\sqrt{q} \).

The implication is that the projection of \( f_0 \) on \( \mathcal{F}_{d}^{LC} \cap \mathcal{F}_d^S \) results in \( \mu_1^*, \ldots, \mu_S^* \) that are arbitrarily close together as \( q \searrow 0 \), i.e. the additional constraint can yield an arbitrarily large approximation error as \( q \searrow 0 \).

Since the approximation error from projecting \( f_n \) on \( \mathcal{F}_d^S \cap \mathcal{F}_{d}^C \) is too large for such a strategy to be sensible, we consider the possibility of projecting \( f_n \) on \( \mathcal{F}_d^S \) directly. Figures 8 and 9, for which we provide theoretical insight in Proposition 6.3, lay rest to this suggestion; theoretical insight for why the same phenomenon does not arise when using a pilot estimator is provided in Proposition 6.5.
Figure 8: Ring example. From top left to bottom right: true ring density; 1000 random draws from the true density; projection of the $1/n$-weighted point masses at the data points on the spherical Gaussian mixture class with 16 components; projection of the $1/n$-weighted point masses at the data points on the spherical Gaussian mixture class with 25 components.

Figure 9: Bivariate gamma example. Left: true gamma density with independent $\Gamma(2, 1)$ marginals; projection of the $1/n$-weighted point masses (based on $n = 250$ observations) at the data points on the spherical Gaussian mixture class with 64 components.

The intuition for Figures 8 and 9 comes from the fact that the criterion function $L_n(f, g)$ is apathetic concerning the ability of a $f \in \mathcal{F}_d^S$ to well approximate $g$ outside the support of $f_n$. We provide further discussion after Propositions 6.5 and 6.3, a proof of which is provided in §8.

Proposition 6.3. Let $f_0 \in \mathcal{F}_d^{LC}$, but $f_0 \notin \mathcal{F}_d^S$ for fixed $S < n$. Write $f_0^* = \psi^*(P; f_0)$ as

$$f_0^*(y) = \sum_{s=1}^{S} \pi_{0,s}^* \phi(y; \mu_{0,s}^*, qI_d)$$

With probability 1, for $g = f_n$ the $\mu_{n,g}^* := \{\mu_{n,g,1}^*, \ldots, \mu_{n,g,S}^*\}$ that minimise the objective function (3.9) for $\pi_0^* = (\pi_{0,1}^*, \ldots, \pi_{0,S}^*) \in \Delta^S$ known (but arbitrary), must satisfy one of the following:

(i) $\mu_{n,g}^* \subset \mathbb{R}^p \setminus \mathcal{R}_{k(n)}^{f_0}$ with probability 1 for all $n > n_0$ ($n_0 < \infty$);
\( \| \mu_{n,g,s}^* \| = \infty \) for all \( s \in \{1, \ldots, S\} \).

In (i), \( R^f_1 \subset R^f_2 \subset \cdots \) is a nested sequence of closed convex sets such that
\[
R^f_\ell := \{ y \in \mathbb{R}^d : f_0(y) \geq r_\ell \} \quad r_\ell > r_k \quad \forall k > \ell
\]
and \( r_k \searrow 0 \) as \( n \to \infty \).

**Remark 6.4.** In the limit as \( n \to \infty \), only scenario (ii) of Proposition 6.3 is possible.

Let \( C_g := \sup_{y \in \mathbb{R}^p} g(y) \). We can associate with any \( r \in [0, C_g) \) a set \( R^g_r := \{ y \in \mathbb{R}^p : g(y) \geq r \} \). Indeed, for \( g \in \mathcal{F}^{LC}_d \) we have a nested sequence of such sets that are closed and convex according to (8.10), and we may associate with any \( r \), an element of the sequence \( \{ R^g_\ell : \ell \in \mathbb{N} \} \) (or otherwise define a further subsequence of sets \( \{ (R^g_\ell)_m : \ell \in \mathbb{N} \} \) containing an element with which \( R^g_r \) coincides).

**Proposition 6.5.** Consider the framework of Proposition 6.3 with \( g = f_0 \) or \( g = \hat{f}^P \) for \( \hat{f}^P \) a pilot estimate obtained using the log-concave maximum likelihood estimator of Cule et al. (2010); \( \hat{f}^P \in \mathcal{F}^{LC}_d \) by construction. The \( \mu_{n,g}^* := \{ \mu_{n,g,1}^*, \ldots, \mu_{n,g,S}^* \} \) that minimise the objective function (3.9) for \( \pi_0^* = (\pi_{0,1}^*, \ldots, \pi_{0,S}^*) \in \Delta^S \) known (but arbitrary), must satisfy one of the following:

(i) \( \sum_{s=1}^S \pi_{0,s}^* \int_{R^g_r} \phi(y; \mu_{n,g,s}^*, qI_d) dy \geq 1/2 \);

(ii) \( \| \mu_{n,g,s}^* \| = \infty \) for all \( s \in \{1, \ldots, S\} \).

In (i) above, \( R^g_r \) is defined by taking the maximum \( r \in [0, C_g) \) such that \( \int_{R^g_r} g(y) dy \geq 1/2 \).

**Remark 6.6.** Scenario (ii) of Proposition 6.5 can be ruled out in practice by initialising the optimisation scheme over a grid of points whose outer edges are dictated by the convex hull of the data.

The intuition for the result in Proposition 6.3 compared with that of Proposition 6.5 does not depend on the log concavity of \( f_0 \), only on the fact that \( f_0 \) and \( \hat{f}^P \) are both Lebesgue densities whilst \( f_n \) is not. Whereas \( \int f_n dP_n = 1 \), \( \int \hat{f}^P dP_n > 1 \) as \( \hat{f}^P \) is normalised to integrate to one in the most meaningful sense. If we were to replace \( f_n \) with a continuous approximation to a mixture of \( n \) spikes at the data points, for instance
\[
f_n(y) = n^{-1} \sum_{i=1}^n \phi(y - Y_i; n^{-1} I)
\]
where \( \phi(y - Y_i; n^{-1} I) \) is the \( d \)-dimensional spherical normal with variance \( 1/n \), we could impose Lebesgue integrability and prevent the mass from piling up at points in \( \mathbb{R}^d \) where no data are observed. However, this continuous approximation is just another example of
a pilot estimator, and one that is not convenient from the standpoint of imposing further constraints.

The need for the pilot estimator in smooth projection estimator is clear from our discussions in this section. Moreover, the comparison to the EM procedure in §5.1 provides empirical evidence that structural constraints exploited in a pilot estimation stage are valuable, despite the inability of the projection procedure to preserve these constraints.

6.2 Discussion of the projection method

We provided in §3.3.3 an iterative least-squares projection algorithm to estimate the mixing parameters \(\pi_1, \ldots, \pi_S\) and the mean vectors \(\mu_1, \ldots, \mu_S\) of the finite mixture representation. This algorithm was designed specifically for the case of the location mixture of spherical Gaussians. We emphasise that the histogram pilot estimator and the projection on the spherical Gaussian mixture class is simply a particular example of our smooth projection procedure for density estimation. As such, the result that \(S \leq n + 1\) is not important for our method. We are interested in providing a succinct representation of a density estimate in multiple and high dimensions, and for this reason, \(S\) should be taken much smaller than \(n + 1\) when the sample size is moderate or large. For the iterative least-squares projection algorithm in §3.3.3, we advocated taking an \(S\) that possesses an integer-valued \(d^{th}\) root. For more general finite mixture classes, we may wish to use an extremely large dictionary of mixture components and to allow the data and the algorithm to select a small number of entries from this dictionary. The supervised nature of our procedure allows us to do this via a simplex LASSO type algorithm, such as that developed and analysed by Koltchinskii (2009).

6.3 Structural constraints

The choice of tuning parameters in density estimation becomes far less critical when we are prepared to assume more about the shape of \(f_0\). Indeed, the log-concavity constraints imposed in Cule et al. (2010) allow the authors to completely circumvent the need to choose any tuning parameter, and when the constraint is satisfied, the shape-constrained maximum likelihood estimator is shown in simulations to deliver improvements over the kernel density estimator when the sample size exceeds some threshold. As a consequence to the omission of tuning parameters, the density estimate that results from the procedure of Cule et al. (2010) is not differentiable; this is highlighted, inter alia, in Delaigle (2010) as a negative feature of the log-concave maximum likelihood estimator. It also forms the motivation for the smoothed log-concave maximum likelihood estimator of Chen and Samworth (2013).

Since we project on a finite dimensional class of densities, which would normally be chosen to possess some smoothness properties, our procedure may be used as a smoothing device
for a non-smooth density estimator. Indeed, the log-concave maximum likelihood estimator could naturally be used as our pilot estimator in order to provide regularisation through its log-concavity constraint.

The histogram is also an estimator for which constraints are easily imposed. In the gamma example that appears in §5.1 we provide two ways to incorporate constraints. Through an independence assumption we are able to construct a histogram by multiplying together the two marginals. Under the assumption that \( f_0 \) is unimodal, the information about the unimodality of the marginals can be easily incorporated into the histogram estimator by simply shifting the edges of the bins such that the number of observations in more extreme bins is no larger than the number of elements in less extreme bins. If \( f_0 \) is the density of an elliptically symmetric random variable, then unimodality constraints can also be incorporated in cases of non-independent marginals by imposing unimodality in the coordinate directions provided by the principal components rather than in the coordinate directions supplied by the standard basis in \( \mathbb{R}^d \).

6.4 Sampling from the estimated density

The density estimate that results from our smooth projection estimator, \( \hat{f}_S \), is particularly easy to sample from as it is just a finite mixture of parametric densities, each of which is readily supported by most standard statistical software packages. For instance, for the example of the finite mixture of Gaussian densities, each component density admits efficient sampling via simple transformations of variates obtained through a linear congruential pseudo random number generator (LGC), implemented in most standard programming languages. Sampling from the log-concave density estimate involves, by contrast, an accept-reject algorithm (see Cule et al., 2010, Appendix B.3), which is computationally more demanding.

7 Appendix A: further propositions

Proposition 7.1 illustrates that there is no loss of generality by choosing \( qI_d \) rather than some other covariance matrix with equal diagonal elements no smaller than \( q \).

Proposition 7.1. Let

\[
\mathcal{F}^\dagger := \left\{ f(\cdot) : f(y) = \int_{\mathcal{M}} \phi(y; \mu, q\Omega^\dagger) dG(\mu), G \in \mathcal{G} \right\}
\]

where \( q\Omega^\dagger \) is an arbitrary covariance matrix whose diagonal elements are all equal to \( q \) for \( q \) fixed in \([q, \bar{q}]\). Then \( \mathcal{F}^\dagger \subseteq \mathcal{F} \).

Proof. Let \( \Sigma \) and \( \Sigma^\dagger \) be two covariance matrices. Consider the function classes

\[
\mathcal{F}_\Sigma := \left\{ f(\cdot) : f(x) = \int \phi(x|\mu, \Sigma) dG(\mu), G \in \mathcal{G} \right\}
\]

25
\[ \mathcal{F}_{\Sigma^\dagger} := \left\{ f(\cdot) : f(x) = \int \phi(x|\mu, \Sigma^\dagger) dG(\mu), \ G \in \mathcal{G} \right\} \]

We will first prove that \( \mathcal{F}_{\Sigma^\dagger} \subseteq \mathcal{F}_{\Sigma} \) as long as \( \Sigma - \Sigma^\dagger \) is non-negative definite.

Suppose that \( \Sigma - \Sigma^\dagger \) is non-negative definite. Let \( X \) be a random variable drawn from an arbitrary \( f \in \mathcal{F} \), then

\[ X \overset{d}= Z + W \quad \text{where} \quad Z \sim N(\mu, \Sigma), \ W \sim G \]

and \( \overset{d}= \) means equality in distribution. Its moment generating function is

\[ E[\exp\{t^T X\}] = E[\exp\{t^T (Z + W)\}] = E[\exp\{t^T Z\}] E[\exp\{t^T W\}] = \exp \left\{ t^T \mu + \frac{1}{2} t^T \Sigma t \right\} E[\exp\{t^T W\}] \]

[by normality of \( Z \) ]

\[ = \exp \left\{ t^T a + \frac{1}{2} t^T \Sigma^\dagger t \right\} \exp \left\{ t^T (\mu - a) + \frac{1}{2} t^T (\Sigma - \Sigma^\dagger) t \right\} E[\exp\{t^T W\}] \]

so

\[ X \overset{d}= Z + W \overset{d}= Z' + W', \]

where

\[ Z' \sim N(a, \Sigma^\dagger), \quad W' \overset{d}= Z'' + W, \quad Z'' \sim N((\mu - a), \Sigma - \Sigma^\dagger) \]

hence \( f \in \mathcal{F}^\dagger \), proving that \( \mathcal{F}^\dagger \subseteq \mathcal{F} \) by the arbitrariness of \( f \in \mathcal{F} \). \( \Sigma - \Sigma^\dagger \) being non-negative definite corresponds to \( \Sigma \succ \Sigma^\dagger \) in the sense of Löwner orderings. It is well known (see e.g. Mosler, 2002) that for any correlation matrix \( \Omega, \ I \succ \Omega \), which proves our claim.

**Proposition 7.2.** There exists a discrete measure \( G_S, \) not necessarily unique such that

\[ \sup_{y \in \mathbb{R}^d} \int_{\mathcal{M}} \phi(y; \mu, qI_d) G_S(d\mu) = \sup_{y \in \mathbb{R}^d} \int_{\mathcal{M}} \phi(y; \mu, qI_d) G_0(d\mu) + O\left(S^{-1/d}\right) \]

under the Prokhorov metric, which metricises the topology of weak convergence.

**Proof.** We first show that the set

\[ \Phi := \{ \phi : \mu \mapsto \phi(y; \mu, qI_d); \mu \in [-M, M]^d \} \]

is a subset of the set \( BL(\mathcal{M}) \) of bounded lipschitz functions on \( \mathcal{M} \),

\[ BL(\mathcal{M}) := \{ f : \mathcal{M} \to \mathbb{R} : \| f \|_{BL} < \infty \} , \]

where \( \| f \|_{BL} = \| f \|_{L} + \| f \|_{\infty} \) with \( \| f \|_{L} = \sup_{x \neq y; x, y \in \mathcal{M}} |f(x) - f(y)|/d(x, y) \).
By the mean value theorem, it suffices to show that the elements of the gradient vector are bounded uniformly over \( y \in \mathbb{R}^d \) and \( \mu \in [-M, M]^d \). To this end, consider
\[
\sup_{y \in \mathbb{R}^d, \mu \in \mathcal{M}} \left| \nabla \mu \phi(y; \mu, q I_d) \right| = \sup_{y \in \mathbb{R}^d, \mu \in \mathcal{M}} \left| (\mu^T I_d)^{1/2} (2\pi^{d/2} \exp\{-q^{-1}(y - \mu)^T(y - \mu)\}) \right|,
\]
which is bounded elementwise since \( q > 0 \) and \( M < \infty \). We have shown that \( \phi : \mu \mapsto \phi(y; \mu, q I_d) \) is in \( BL(\mathcal{M}) \).

Now suppose for a contradiction that no discrete measure \( G_S \) exists such that
\[
\sup_{y \in \mathbb{R}^d} \int_\mathcal{M} \phi(y; \mu, q I_d) G_S(d\mu) \to \sup_{y \in \mathbb{R}^d} \int_\mathcal{M} \phi(y; \mu, q I_d) G_0(d\mu) \quad \text{as} \quad S \to \infty.
\]
Introduce the set of discrete measures
\[
G^\epsilon_S := \left\{ G_S^\epsilon : G_S^\epsilon(A) = \sum_{s=1}^S \pi_s \delta_{\mu_s}^\epsilon(A) : \pi_1, \ldots, \pi_S \geq 0; \sum_{s=1}^S \pi_s = 1; A \in \mathcal{B}(\mathcal{M}) \right\}
\]
where \( \mathcal{B}(\mathcal{M}) \) is the Borel sigma-algebra on \( \mathcal{M} \),
\[
\delta_{\mu_s}^\epsilon(A) = \begin{cases} 1 & \text{if } \mu_s \in A \\ 0 & \text{otherwise}, \end{cases}
\]
and \( \mu_1^\epsilon, \ldots, \mu_S^\epsilon \) are \( S \) elements of \( \mathcal{M} = [-M, M]^d \) satisfying \( \cup_{s=1}^S B(\mu_s^\epsilon, \epsilon) = \mathcal{M} \) and \( B(\mu_s^\epsilon, \epsilon) \cap B(\mu_j^\epsilon, \epsilon) = 0 \) for all \( s \neq j, s, j \in \{1, \ldots, S\} \). i.e. \( \{B(\mu_s^\epsilon, \epsilon) : s = 1, \ldots, S\} \) are disjoint open balls of radius \( \epsilon/2 \) around \( \{\mu_s^\epsilon : s = 1, \ldots, S\} \) whose union covers \( \mathcal{M} \). It is easy to see that such a collection of \( \{\mu_s^\epsilon : s = 1, \ldots, S\} \) exists since \( \mathcal{M} \) is compact.

Let \( \mathcal{G}(\mathcal{M}) \) be the set of all probability measures on the Borel sets of \( \mathcal{M} \). Since \( G_S^\epsilon \) is dense in \( \mathcal{G}(\mathcal{M}) \) under the weak topology (Parthasarathy, 1967, Theorem 6.3), there exists a choice of weights \( \{\pi_s : s = 1, \ldots, S\} \) such that the sequence of weighted discrete measures converges to \( G_0 \) as \( S \to \infty \). It follows by Theorem 11.3.3 of Dudley (2002) that
\[
\int f dG^\epsilon_s \to \int f dG_0 \quad \text{for all} \quad f \in BL(\mathcal{M}),
\]

hence it cannot be true that \( \phi : \mu \mapsto \phi(y; \mu, q I_d) \) is in \( BL(\mathcal{M}) \), a contradiction.

We next prove that \( \rho(G_S^\epsilon, G_0) = O(S^{-d/2}) \) where \( \rho(\cdot, \cdot) \) is the Prohorov metric defined below.

**Definition 7.3.** Let \( \mathbb{Q} \) and \( \mathbb{P} \) be laws on \( \mathcal{M} \). The Prohorov metric is
\[
\rho(\mathbb{Q}, \mathbb{P}) := \inf \{\epsilon > 0 : \mathbb{Q}(A) \leq \mathbb{P}(A^\epsilon) + \epsilon \quad \text{for all Borel sets} \quad A\}
\]
where \( A^\epsilon := \{y \in \mathcal{M} : d(x, y) < \epsilon \text{ for some } x \in A\} \), i.e. the “\( \epsilon \)-enlargement” of \( A \).
Fix $\epsilon > 0$ and introduce the function

$$h(\mu) = 0 \vee (1 - d(\mu, A)/\epsilon)$$

where $d(\mu, A) = \inf_{r \in A} d(\mu, r)$. Then $h \in BL(M)$ (Dudley, 2002, page 396 and Proposition 11.2.2) and $\mathbb{I}\{\mu \in A\} \leq h(\mu) \leq \mathbb{I}\{\mu \in A^c\}$. Cover $[-M, M]^d$ with disjoint open balls or radius $\epsilon/2$ around $\{\mu_s : s = 1, \ldots, S\}$ and fix weights $\{\pi_s : s = 1, \ldots, S\}$ such that

$$\sum_{s=1}^{S} |\pi_s \delta_{\mu_s} - G_0(B^s_\epsilon)| \leq \epsilon,$$

where $B^s_\epsilon := B(\mu^s_\epsilon, \epsilon)$. For an arbitrary Borel set, $A$,

$$G^e_S(A) \leq \int_M h(\mu)G_0(d\mu) + \int_M h(\mu)|G^e_S - G_0|(d\mu)$$

$$\leq \int_M \mathbb{I}\{\mu \in A^c\}G_0(d\mu) + \int_M h(\mu)|G^e_S - G_0|(d\mu)$$

$$\leq G_0(A^c) + \sup_{r \in M} |h(r)| \sum_{s=1}^{S} |\pi_s \delta_{\mu_s} - G_0(B^s_\epsilon)| + \sup_{r \in M} |h(r)|G_0((\cup_{s=1}^{S} B^s_\epsilon)^c)$$

$$= G_0(A^c) + \epsilon.$$

Since we are allowed $S$ balls to cover $M = [-M, M]^d$, the minimum $\epsilon$ for which the last line of the above display holds is $\epsilon = 2M/S^{1/d}$, hence the Prohorov metric converges at rate $S^{-1/d}$ when $G_S$ is taken as $G^e_S$.

\[ \square \]

8 Appendix B: proofs of main results

We collect below the longer proofs for the theoretical results appearing in §4 and §6.

Proof. [Theorem 4.4] Decompose as

$$\left\| \psi^*(P_n, \hat{f}^P) - \psi^*(P, f_0) \right\|_2 \leq \left\| \psi^*(P_n, \hat{f}^P) - \psi^*(P, \hat{f}^P) \right\|_2 + \left\| \psi^*(P, \hat{f}^P) - \psi^*(P, f_0) \right\|_2$$

$$= I + II$$

Control over I

Let $\ell_\theta$ denote the loss function $\ell_\theta = (\hat{f}^P - f)^2$. Then we may write $\int (\hat{f}^P - f)^2 dP_n$ and $\int (\hat{f}^P - f)^2 dP$ as $\mathbb{M}_n = P_n \circ \ell_\theta$ and $\mathbb{M} = P \circ \ell_\theta$ respectively. $\mathbb{M}_n$ and $\mathbb{M}$ are stochastic processes indexed by the metric space $L_1(\Theta)$ where $\Theta \subset \mathbb{R}^L$, $L \geq S$. Viewing $\hat{f}^P$ as a non-random object, the uniform weak large of large numbers together with Slutsky’s Theorem guarantee that $\sup_{\theta \in \Theta} |\mathbb{M}_n(\theta) - \mathbb{M}(\theta)| \rightarrow_p 0$. 28
Since there exists a $\theta^*_p$ such that $f(\theta^*_p)$ is the unique minimiser of $\int (\hat{f}^P - f)^2 dP$ over $\mathcal{F}^S_d$,
\[
\mathbb{M}(\theta^*_p) < \inf_{\theta \in \Theta} \mathbb{M}(\theta)
\]
for every open subset that contains $\theta^*_p$, therefore we can construct a sequence $\theta^p_1, \ldots, \theta^p_n$ such that
\[
\mathbb{M}_n(\theta^p_n) \leq \inf_{\theta \in \Theta} \mathbb{M}_n(\theta) + o_p(1).
\]
By Corollary 3.2.3 of van der Vaart and Wellner (1996), $\theta^p_n \rightarrow_p \theta^*_p$. Moreover, by the definition of $f(\theta^*_p)$ as a minimiser, it follows through a Taylor expansion of $\mathbb{M}(\theta^*_p) - \mathbb{M}(\theta)$ around $\theta^*_p$ that
\[
\mathbb{M}(\theta^p) - \mathbb{M}(\theta) \lesssim -\|\theta^*_p - \theta\|^2,
\]
thus $\|\ell_{\theta^*_p} - \ell_{\theta^p}\| = O_p(n^{-1/2})$ by Theorem 3.2.5 of van der Vaart and Wellner (1996), and correspondingly $\|f(\theta^*_p) - f(\theta^p)\| = O_p(n^{-1/2})$.

Control over II

Introduce the functional $\Phi(g; f) = \int_{\mathbb{R}^d} (g - f)^2 dP$. We will make use of the following definition of the Fréchet derivative.

**Definition 8.1.** For normed spaces $\mathbb{D}_\Phi$ and $\mathbb{E}$ and for some map $\Phi : \mathbb{D}_\Phi \mapsto \mathbb{E}$, the Fréchet derivative (if it exists) is the linear continuous map $D\Phi_g : \mathbb{D} \mapsto \mathbb{E}$ such that
\[
\|\Phi(g + h) - \Phi(g) - D\Phi_g(h)\| = o(||h||).
\]

We have
\[
D\Phi_g(h; f) = 2 \int_{\mathbb{R}^d} (h - f)hdP
\]
with
\[
\|\Phi(g + h; f) - \Phi(g; f) - D\Phi_g(h; f)\| = o(||h||).
\]
Letting $h = \hat{f}^P - f_0$ and $g = f_0$,
\[
\left\| \Phi(\hat{f}^P; f) - \Phi(f_0; f) - D\Phi_g(\hat{f}^P - f_0; f) \right\| = 2 \left\| \int_{\mathbb{R}^d} (\hat{f}^P - f)^2 dP - \int_{\mathbb{R}^d} (f_0 - f)^2 dP - 2 \int_{\mathbb{R}^d} \left( (\hat{f}^P - f_0) - f \right)(\hat{f}^P - f_0) \right\| = o(||\hat{f} - f_0||^2)
\]
Since $\hat{f}^P$ satisfies $r_n(\hat{f}^P - f_0) \rightsquigarrow X$, it follows by the functional delta method (van der Vaart and Wellner, 1996, Theorem 3.9.5) that
\[
r_n \left( \Phi(\hat{f}^P; f) - \Phi(f_0; f) \right) \rightsquigarrow D\Phi_g(X),
\]
thus
\[
\|\Phi(\hat{f}^P; f) - \Phi(f_0; f)\|_2 = o(r_n^{-1}).
\]
Since $\Phi(\cdot; f)$ is continuous in $f$, an application of Slutsky’s theorem completes the proof. \qed
Proof. [Proposition 6.2] We will show that
\[ \sum_{s=1}^{S} \phi(y; \mu_s, q) \| \rho_s \| \leq 1 \quad \forall \ y \in \text{supp}(f) \]
is the only scenario under which the Hessian matrix of the log density \( f \) (for \( f \in \bar{F}_d^S \)) is guaranteed to be negative semi-definite. Let \( f_s \) denote the \( s \)th mixture component of \( f \). The gradient vector of \( f \) is
\[
g := \frac{\partial f}{\partial y} = \sum_{s=1}^{S} \pi_s f_s \frac{\partial f_s}{\partial y} = \sum_{s=1}^{S} \pi_s f_s q^{-1} I_d (y - \mu_s) \]
and its Hessian is
\[
H := \frac{\partial^2 f}{\partial y \partial y^T} = \sum_{s=1}^{S} \pi_s f_s q^{-1} I_d ((y - \mu_s)(y - \mu_s)^T - q I_d) q^{-1} I_d. \]
The Hessian of \( \log f \) is
\[
\frac{\partial^2 \log f}{\partial y \partial y^T} = -\frac{1}{f^2} gg^T + \frac{1}{f} H = \frac{1}{q f^2} \left( \sum_{s=1}^{S} \pi_s f_s \frac{(y - \mu_s)}{q^{1/2}} \right) \left( \sum_{s=1}^{S} \pi_s f_s \frac{(y - \mu_s)}{q^{1/2}} \right)^T + \frac{1}{q f} \sum_{s=1}^{S} \pi_s f_s \left( \frac{(y - \mu_s)}{q^{1/2}} \right) \left( \frac{(y - \mu_s)}{q^{1/2}} \right)^T - q I_d. \]
The first term is always negative semi-definite. Let \( \rho_s = (y - \mu_s)/q^{1/2} \). As shown in the proof of Theorem D.4 of Carreira-Perpinan (2000), all eigenvalues of
\[
\sum_{s=1}^{S} \pi_s f_s \rho_s \rho_s^T - I \]
are bounded above by \(-1 + \sum_{s=1}^{S} f_s \| \rho_s \|^2 \), hence the second term is negative semi-definite if \( \sum_{s=1}^{S} f_s \| \rho_s \|^2 \leq 1 \). \( \square \)

Proof. [Proposition 6.3] Since \( \pi_0^* = (\pi_{0,1}^*, \ldots, \pi_{0,S}^*) \) are known, the projection \( f_{n,g}^* = f(\mu_{n,g,1}^*, \ldots, \mu_{n,g,S}^*, \pi_{0}^*) \) of \( g \) on \( \bar{F}_d^S \) satisfies
\[
0 = \frac{\partial L_n(f(\pi_0, \mu); g)}{\partial \mu_s} (\mu_{n,g}^*, \pi_{0}^*) \forall s \in \{1, \ldots, S\}, \quad \mu_{n,g}^* = (\mu_{n,g,1}^*, \ldots, \mu_{n,g,S}^*), \quad (8.1) \]
i.e.
\[ 0 = 2 \sum_{i=1}^{n} (g(Y_i) - \sum_{s=1}^{S} \pi_{0,s}^{*} \phi(Y_i; \mu_{n,g,s}^{*}, q I_d)) V_s(Y_i; \pi_{0,s}^{*}, \mu_{n,g,s}^{*}) \quad \forall s \in \{1, \ldots, S\} \quad (8.2) \]

where
\[ V_s(Y_i; \pi_{0,s}^{*}, \mu_{n,g,s}^{*}) = \pi_{0,s}^{*} \phi(Y_i; \mu_{n,g,s}^{*}, q I_d) q^{-1} I_d (Y_i - \mu_{n,g,s}^{*}). \quad (8.3) \]

From (8.2) we see that, at a minimising \( \{\mu_{n,g,1}^{*}, \ldots, \mu_{n,g,S}^{*}\} \), \( \{Y_1, \ldots, Y_n\} \subset \mathcal{J} \cup \mathcal{E} \), where \( \mathcal{I} = \mathcal{J} \cup \mathcal{E} \),
\[ \mathcal{J}^c := \left\{ Y_i : \sum_{j : Y_j \in \mathcal{E}} \left( g(Y_i) - \sum_{s=1}^{S} \pi_{0,s}^{*} \phi(Y_i; \mu_{n,g,s}^{*}, q^{-1} I_d) \right) = 0 \right\}. \quad (8.4) \]
\[ \mathcal{J} := \left\{ Y_i : V_s(Y_i; \pi_{0,s}^{*}, \mu_{n,g,s}^{*}) = 0 \quad \forall s \in \{1, \ldots, S\} \right\}. \quad (8.5) \]

and
\[ \mathcal{E} := \left\{ Y_i : \sum_{j : Y_j \in \mathcal{E}} \left( g(Y_i) - \sum_{s=1}^{S} \pi_{0,s}^{*} \phi(Y_i; \mu_{n,g,s}^{*}, q^{-1} I_d) \right) = 0 \right\}. \quad (8.6) \]

Notice that \( Y_i \in \mathcal{J} \) if and only if \( ||Y_i - \mu_{n,g,s}^{*}|| = \infty \) for all \( s \in \{1, \ldots, S\} \), thus if there exists a \( Y_i \) in \( \mathcal{J} \), then \( \mathcal{J} = \{Y_1, \ldots, Y_n\} \); this is scenario (ii). We have \( \mathcal{J} = \{Y_1, \ldots, Y_n\} \) or \( \mathcal{J} = \emptyset \).

Consider \( \mathcal{J} = \emptyset \), then \( \mathcal{J}^c = \mathcal{I} \cup \mathcal{E} \). Since \( S < n \), there exists a set \( \mathcal{S} \neq \emptyset \) such that \( Y_i \notin \mathcal{I} \) for all \( Y_i \in \mathcal{S} \). Suppose for a contradiction that \( \mu_{n,g}^{*} := \{\mu_{n,g,1}^{*}, \ldots, \mu_{n,g,1}^{*}\} \subset R_{k(n)}^{f_0} \) yields a minimum, then \( \mathcal{S} = \mathcal{E} \) and therefore \( \mathcal{E} \neq \emptyset \).

Introduce the sets
\[ \mathcal{E}_g^{-} := \left\{ Y_i : \sum_{s=1}^{S} \pi_{0,s}^{*} \phi(Y_i; \mu_{n,g,s}^{*}, q^{-1} I_d) < w_{g,i} \right\} \quad (8.7) \]
and
\[ \mathcal{E}_g^{+} := \left\{ Y_i : \sum_{s=1}^{S} \pi_{0,s}^{*} \phi(Y_i; \mu_{n,g,s}^{*}, q^{-1} I_d) > w_{g,i} \right\}. \quad (8.8) \]

and notice that at a minimum \( \mathcal{E} = \mathcal{E}_g^{+} \cup \mathcal{E}_g^{-} \), with \( \mathcal{E}_g^{-} \neq \emptyset \) and \( \mathcal{E}_g^{+} \neq \emptyset \) by the definition of \( \mathcal{E} \) and the fact that \( \mathcal{E} \neq \emptyset \). Re-writing the clause in \( \mathcal{E} \) more explicitly in terms of \( \mathcal{E}_g^{-} \) and \( \mathcal{E}_g^{+} \), we have
\[ |\mathcal{E}_g^{-}| \left( \frac{1}{|\mathcal{E}_g^{-}|} \sum_{i : Y_i \in \mathcal{E}_g^{-}} \left( w_{g,i} - \sum_{s=1}^{S} \pi_{0,s}^{*} \phi(Y_i; \mu_{n,g,s}^{*}, q I_d) \right) \right) \]
\[ = -|\mathcal{E}_g^{+}| \left( \frac{1}{|\mathcal{E}_g^{+}|} \sum_{i : Y_i \in \mathcal{E}_g^{+}} \left( w_{g,i} - \sum_{s=1}^{S} \pi_{0,s}^{*} \phi(Y_i; \mu_{n,g,s}^{*}, q I_d) \right) \right). \quad (8.9) \]
By log concavity of $f_0$ and $\widehat{f}^P$, we can define a nested sequence of closed convex sets $\mathcal{R}^0_1 \subset \mathcal{R}^0_2 \subset \cdots$ such that

$$\mathcal{R}^0_\ell := \{ y \in \mathbb{R}^d : f_0(y) \geq r_\ell \} \quad r_\ell > r_k \ \forall k > \ell. \quad (8.10)$$

We can similarly define

$$\mathcal{R}^0_{n,\ell} := \{ y \in \{Y_1, \ldots, Y_n\} : f_0(y) \geq r_\ell \} \quad r_\ell > r_k \ \forall k > \ell. \quad (8.11)$$

We will show that $\mu^\ast_{n,g} \subset \mathcal{R}^0_{n,k(n)}$ yields a contradiction with (8.9) when $f_0 \in \mathcal{F}^{LC}_d$. Since $r_n \searrow 0$ as $n \to \infty$, we can write

$$\lim_{n \to \infty} \mathcal{R}^0_{k(n)} = \mathcal{E} := \{ y \in \mathbb{R}^p : f_0 \geq 0 \},$$

i.e. $\mathcal{E} = \text{supp}(f_0)$. Let $\mathcal{R}^0_{n,k(n)}$ denote the empirical analogue of $\mathcal{R}^0_{k(n)}$ as defined by equation (8.11), and likewise for $\mathcal{E}^\ast_{n}$. Since $f_0 \in \mathcal{F}^{LC}_d$, there exists an $n'$ such that for all $n > n'$,

$$|\mathcal{R}^0_{n,k(n)}| > |\mathcal{E}^\ast_{n} \setminus \mathcal{R}^0_{n,k(n)}| \quad \text{and} \quad |\mathcal{R}^0_{n,k(n)} \cap \mathcal{E}^\ast_{n}| = |\mathcal{R}^0_{n,k(n)}|, \quad \text{both with probability } 1.$$ 

Since $|\mathcal{R}^0_{n,k(n)}| + |\mathcal{E}^\ast_{n} \setminus \mathcal{R}^0_{n,k(n)}| = n$, we conclude that $|\mathcal{E}^\ast_{n}| > |\mathcal{E}^\ast_{n}|$ for all $n > n'$ with probability 1. Moreover, since $w_{g,i} = n^{-1}$,

$$\frac{1}{|\mathcal{E}^\ast_{g}|} \sum_{i : Y_i \in \mathcal{E}^\ast_{g}} \left( w_{g,i} - \sum_{s=1}^{S} \pi_{0,s}^\ast \phi(Y_i; \mu_{n,g,s}^\ast, qI_d) \right) < \frac{|\mathcal{E}^\ast_{g}|}{|\mathcal{E}^\ast_{g}|} \left( \frac{1}{n} \right)$$

whilst

$$-\frac{1}{|\mathcal{E}^\ast_{g}|} \sum_{i : Y_i \in \mathcal{E}^\ast_{g}} \left( w_{g,i} - \sum_{s=1}^{S} \pi_{0,s}^\ast \phi(Y_i; \mu_{n,g,s}^\ast, qI_d) \right) \nearrow C > 0.$$ 

Therefore there exists an $n_0 \geq n'$ such that for all $n > n_0$

$$-|\mathcal{E}^\ast_{g}| \left( \frac{1}{|\mathcal{E}^\ast_{g}|} \sum_{i : Y_i \in \mathcal{E}^\ast_{g}} \left( w_{g,i} - \sum_{s=1}^{S} \pi_{0,s}^\ast \phi(Y_i; \mu_{n,g,s}^\ast, qI_d) \right) \right)$$

$$\geq \frac{|\mathcal{E}^\ast_{g}|}{|\mathcal{E}^\ast_{g}|} \sum_{i : Y_i \in \mathcal{E}^\ast_{g}} \left( w_{g,i} - \sum_{s=1}^{S} \pi_{0,s}^\ast \phi(Y_i; \mu_{n,g,s}^\ast, qI_d) \right).$$

This proves that there exists at least one $Y_i \in \mathcal{S}$ such that $Y_i \notin \mathcal{E}$, which is a contradiction to $\mathcal{S} = \mathcal{E}$. We conclude that a minimum is unobtainable with $\mathcal{J} = \emptyset$ and $\mu_{n,g}^\ast \subset \mathcal{R}^0_{n,k(n)}$; if $\mathcal{J} = \emptyset$ at a minimum, then $\mu_{n,g}^\ast \subset \mathbb{R}^p \setminus \mathcal{R}^0_{n,k(n)}$. \qed

**Proof.** [Proposition 6.5] The sets $\mathcal{J}$, $\mathcal{I}^c$ and $\mathcal{E}$ are those of equations (8.5), (8.4) and (8.6) respectively. As in the proof of Proposition 6.5, since $f_0 \in \mathcal{F}^{LC}_d$, either $\mathcal{J} = \{Y_1, \ldots, Y_n\}$

32
or $\mathcal{J} = \emptyset$. If $\mathcal{J} = \{Y_1, \ldots, Y_n\}$, then the minimum is the one with $||\mu_{n,g,s}|| = \infty$ for all $s \in \{1, \ldots, S\}$.

Consider minima achieved with $\mathcal{J} = \emptyset$. Then for any $Y_i \in \{Y_1, \ldots, Y_n\}$, $Y_i \in \mathcal{J}^c = \mathcal{I}^c \cup \mathcal{E}$ at a minimum. Since $f_0 \notin \mathcal{F}_d^S$ and $\widehat{f}_P \notin \mathcal{F}_d^S$, with probability 1 there exists a $S \neq \emptyset$ such that $Y_i \notin \mathcal{I}^c$ for all $Y_i \in S$. Hence $S \neq \emptyset$ with probability 1, and at a minimum $S = \mathcal{E}$.

As in the proof of Proposition 6.3, introduce the sets $\mathcal{E}_g^-$ and $\mathcal{E}_g^+$ (see equations (8.7) and (8.8) respectively) and notice that at a minimum $\mathcal{E} = \mathcal{E}_g^+ \cup \mathcal{E}_g^-$, with $\mathcal{E}_g^- \neq \emptyset$ and $\mathcal{E}_g^+ \neq \emptyset$ by the definition of $\mathcal{E}$ and the fact that $\mathcal{E} \neq \emptyset$.

Clearly, when $\mu_{n,g}^\ast$ violates condition (i),

\[
\sum_{i: Y_i \in \mathcal{E}_g^-} \left( w_{g,i} - \sum_{s=1}^{S} \pi_{0,s}^\ast \phi(Y_i; \mu_{n,g,s}^\ast, q I_d) \right) > - \sum_{i: Y_i \in \mathcal{E}_g^+} \left( w_{g,i} - \sum_{s=1}^{S} \pi_{0,s}^\ast \phi(Y_i; \mu_{n,g,s}^\ast, q I_d) \right)
\]

with probability 1 for sufficiently large $n$, thereby contradicting the requirement for a minimum that $Y_i \in \mathcal{E}$ whenever $Y_i \notin \mathcal{I}^c$.

\[\Box\]

**Acknowledgements:** The authors thank Matthew Arnold, John Aston and Richard Samworth, as well as seminar attendees at Bath, Paris VI, Princeton, Sheffield and Warwick for their comments on aspects of this work. The first author is grateful for a visiting research fellowship from Princeton University and financial support from the Engineering and Physical Sciences Research Council under grants EP/D063485/1 and EP/K004581/1.

**References**

Augustin, N., Trenkel, V., Wood, S. and Lorance, P. (2013). Space-time modelling of blue ling for fisheries stock management. *Environmetrics* In press.

Balabdaoui, F. and Wellner, J. A. (2010). Estimation of a $k$-monotone density: characterizations, consistency and minimax lower bounds. *Stat. Neerl.* 64 45–70.

Bickel, P. J. and Levina, E. (2008). Covariance regularization by thresholding. *Ann. Statist.* 36 2577–2604.

Braun, W. J. and Hall, P. (2001). Data sharpening for nonparametric inference subject to constraints. *J. Comput. Graph. Statist.* 10 786–806.
CARREIRA-PERPINAN, M. A. (2000). Mode-finding for mixtures of gaussian distributions. *IEEE Trans. on Pattern Analysis and Machine Intelligence* **22** 1318–1323.

CHEN, Y. and SAMWORTH, R. (2013). Smoothed log-concave maximum likelihood estimation with applications. *Statist. Sinica* .

CULE, M., SAMWORTH, R. and STEWART, M. (2010). Maximum likelihood estimation of a multidimensional log-concave density. *J. R. Stat. Soc. Ser. B Stat. Methodol.* With discussion.

DELAIGLE, A. (2010). Discussion of *Maximum likelihood estimation of a multi-dimensional log-concave density*. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **72** 545–607.

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39** 1–38. With discussion.

DEVROYE, L., GYÖRFI, L. and LUGOSI, G. (1996). *A probabilistic theory of pattern recognition*, vol. 31 of *Applications of Mathematics* (New York). Springer-Verlag, New York.

DEVROYE, L. and LUGOSI, G. (2004). Bin width selection in multivariate histograms by the combinatorial method. *Test* **13** 129–145.

DUDLEY, R. M. (2002). *Real analysis and probability*, vol. 74 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge. Revised reprint of the 1989 original.

DUONG, T. (2007). *Package ‘ks’: Kernel smoothing*. R package version 1.4.11. URL [http://cran.r-project.org/web/packages/ks/index.html](http://cran.r-project.org/web/packages/ks/index.html)

DUONG, T. and HAZELTON, M. L. (2003). Plug-in bandwidth matrices for bivariate kernel density estimation. *J. Nonparametr. Stat.* **15** 17–30.

EL KAROUI, N. (2008). Operator norm consistent estimation of large-dimensional sparse covariance matrices. *Ann. Statist.* **36** 2717–2756.

FIX, E. and HODGES, J. (1951). Discriminatory analysis nonparametric discrimination: Consistency properties. *Technical report, USAF School of Aviation Medicine* .

HALL, P. and PRESNELL, B. (1999). Biased bootstrap methods for reducing the effects of contamination. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **61** 661–680.

JOHNSTONE, I. M. and LU, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *J. Amer. Statist. Assoc.* **104** 682–693.
KANAZAWA, Y. (1992). An optimal variable cell histogram based on the sample spacings. Ann. Statist. 20 291–304.

KOLTCHINSKII, V. (2009). Sparse recovery in convex hulls via entropy penalisation. Ann. Statist. 37 1332–1359.

LAM, C. and FAN, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. Ann. Statist. 37 4254–4278.

LIEBSCHER, E. (2005). A semiparametric density estimator based on elliptical distributions. J. Multivariate Anal. 92 205–225.

LIU, H., HAN, F., YUAN, M., LAFFERTY, J. and WASSERMAN, L. (2012a). High dimensional semiparametric gaussian copula graphical models. Ann. Stat. 40 2293–2326.

LIU, H., HAN, F. and ZHANG, C.-H. (2012b). High dimensional transelliptical graphical models. Neural Information Processing Systems (NIPS) 25.

LIU, H., LAFFERTY, J. and WASSERMAN, L. (2007). Sparse nonparametric density estimation in high dimensions using the rodeo. Journal of Machine Learning Research 2 283–290.

LIU, H., LAFFERTY, J. and WASSERMAN, L. (2009). The nonparanormal: semiparametric estimation of high dimensional undirected graphs. J. Mach. Learn. Res. 10 2295–2328.

LIU, H., XU, M., GU, H., GUPTA, A., LAFFERTY, J. and WASSERMAN, L. (2011). Forest density estimation. J. Mach. Learn. Res. 12 907–951.

MOSLER, K. (2002). Multivariate dispersion, central regions and depth, vol. 165 of Lecture Notes in Statistics. Springer-Verlag, Berlin. The lift zonoid approach.

PARTHASARATHY, K. R. (1967). Probability measures on metric spaces. Probability and Mathematical Statistics, No. 3, Academic Press Inc., New York.

RICHARDSON, S. and GREEN, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components. J. Roy. Statist. Soc. Ser. B 59 731–792.

ROCKAFELLAR, R. T. (1997). Convex analysis. Princeton Landmarks in Mathematics, Princeton University Press, Princeton, NJ. Reprint of the 1970 original, Princeton Paperbacks.

SAMWORTH, R. J. and YUAN, M. (2012). Independent component analysis via nonparametric maximum likelihood estimation. Ann. Stat. 40 2973–3002.
Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. J. Roy. Statist. Soc. Ser. B 53 683–690.

Stone, C. J. (1980). Optimal rates of convergence for nonparametric estimators. Ann. Statist. 8 1348–1360.

Stone, C. J. (1985). An asymptotically optimal histogram selection rule. In Proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer, Vol. II (Berkeley, Calif., 1983). Wadsworth Statist./Probab. Ser., Wadsworth, Belmont, CA.

Street, W. N., Wolberg, W. H. and Mangasarian, O. L. (1993). Nuclear feature extraction for breast tumor diagnosis. Proc. SPIE 1905.

van der Vaart, A. W. and Wellner, J. A. (1996). Weak convergence and empirical processes. Springer Series in Statistics, Springer-Verlag, New York. With applications to statistics.

Wand, M. P. and Jones, M. C. (1995). Kernel smoothing, vol. 60 of Monographs on Statistics and Applied Probability. Chapman and Hall Ltd., London.

Wood, S. N., Bravington, M. V. and Hedley, S. L. (2008). Soap film smoothing. J. R. Stat. Soc. Ser. B Stat. Methodol. 70 931–955.

Yuan, M. (2009). State price density estimation via nonparametric mixtures. Ann. Appl. Stat. 3 963–984.

Zou, H., Hastie, T. and Tibshirani, R. (2006). Sparse principal component analysis. J. Comput. Graph. Statist. 15 265–286.