A Geometric Approach to Density Estimation with Additive Noise

Stefan Wager
Department of Statistics, Stanford University
Stanford, CA-94305, U.S.A.
swager@stanford.edu

February 19, 2013

Abstract

We introduce and study a method for density estimation under an additive noise model. Our method does not attempt to maximize a likelihood, but rather is purely geometric: Heuristically, we L_2-project the observed empirical distribution onto the space of candidate densities that are reachable under the additive noise model. Our estimator reduces to a quadratic program, and so can be computed efficiently. In simulation studies, it roughly matches the accuracy of fully general maximum likelihood estimators at a fraction of the computational cost. We also give a theoretical analysis of the estimator, and show that it is consistent, attains a quasi-parametric convergence rate under moment conditions, and is robust to model mis-specification. We provide an R implementation of the proposed estimator in the package nlpden.

Keywords: M-estimator, minimum distance estimator, mixture model, quadratic program, shape constrained estimator.

1 Introduction

Consider the high-dimensional Gaussian noise model, in which we observe

$$X_i = \mu_i + \varepsilon_i,\ \varepsilon_i \sim \mathcal{N}(0,1) \text{ independently for } 1 \leq i \leq n. \tag{1}$$

Following Robbins [1964], this model has often been analyzed from an empirical Bayes perspective. In a classical Bayesian setting, we assume that $\mu_i \sim G$ for some prior distribution G with density g (or, more generally, Radon-Nikodym derivative g) and that the observations X_i are distributed according to the convolution density $f = \varphi * g$, where φ is the standard normal density. The challenge in an empirical Bayes setting is that g and f are unknown, and must be estimated from the X_i.

Having a good estimate $\hat{f}$ for the marginal density f of the X_i is useful. For example, $\hat{f}$ lets us estimate posterior means $\hat{\mu}_i = \mathbb{E}[\mu_i|X_i]$ [e.g. recently Brown and Greenshtein, 2009, Jiang and Zhang, 2009]. Johnstone and Silverman [2004] showed this formalism to be useful for sparse signal detection, and Efron [2011] suggested it as a cure for selection bias. Moreover, the fitted density $\hat{f}$ gives us a natural estimate $\varphi(X_i)/\hat{f}(X_i)$ for the local false discovery rate [Efron et al., 2001], which is a useful upper bound on the posterior probability that the i^{th} effect μ_i is zero.

At first glance, we might expect the problem of estimating f to be fairly straightforward, provided we accept the model (1). The family of densities that can be written as $f = \varphi * g$ is fairly small, suggesting that selecting the density $\hat{f} = \varphi * \hat{g}$ that is "closest" to the empirical distribution of the X_i should be a simple and well-behaved operation.

Most papers aiming to use the Gaussian noise model in an empirical Bayes analysis, however, use general purpose density estimation techniques rather than the special form of (1) in producing an estimate $\hat{f}$ for the density of the X_i. Kernel smoothing methods are quite popular for estimating functionals of the μ_i [e.g. Brown and Greenshtein, 2009, Butucea and Comte, 2009, Jiang, 2012, Zhang, 1997]. In the local false discovery rate literature, Efron [2007, 2010] has advocated the use of either Poisson regression or log-splines for estimating f (note that these methods can also be used to estimate false discovery rates when the structural model (1) does not hold). The papers that do explicitly use the model (1) tend to impose stringent constraints on the form of g: Johnstone and Silverman [2004] assume that g is the mixture of a point mass at 0 and a Laplace (or quasi-Cauchy) density, while Muralidharan [2010] models g as a mixture of a finite number of Gaussian bumps $\mathcal{N}(\mu_j, \sigma_j^2)$.

Jiang and Zhang [2009] and Zhang [2009] have developed general maximum likelihood techniques for estimating f, using ideas that go back to e.g., Laird [1978]. The results achieved by these techniques on simulated data are impressive, but the algorithms used to compute the general maximum likelihood estimate for f are still computationally intensive and sensitive to initialization. That being said, recent advances such as the interior point formulation of Koenker and Mizera [2012] should make general maximum likelihood density estimation more computationally tractable in the future.

This paper introduces a simple non-parametric method for estimating f that takes advantage of the assumption that $f = \varphi * g$ for some probability distribution G. In contrast to most currently available estimators that make explicit use of the Gaussian convolution model, our estimator is not motivated by likelihood-based arguments. Rather, our approach is purely geometric: We estimate f using $\hat{f}$, where $\hat{f}$ is the closest density to the empirical distribution of the X_i under L_2 norm such that $\hat{f} = \varphi * \hat{g}$.

More precisely, let $Q(\mathbb{R})$ denote the space of sub-probability distributions (i.e. distributions whose total mass is ≤ 1), let $L_2(\mathbb{R})$ denote the space of square-integrable real-valued functions, and define

$$\mathcal{D}(\mathbb{R}) = \left\{ g : g \geq 0,\ G(a) = \int_{-\infty}^{a} g(dx) \in Q(\mathbb{R}) \right\} \tag{2}$$

Then, the closure under L_2 norm of the space of acceptable marginal densities in the convolution model can be written as

$$\mathcal{E}_\varphi = \left\{ f \in L_2(\mathbb{R}) : f = \varphi * g,\ g \in \mathcal{D}(\mathbb{R}) \right\}. \tag{3}$$

We define a projection operator

$$\begin{aligned}&P_\varphi : \mathcal{D}(\mathbb{R}) \to \mathcal{E}_\varphi \\ &P_\varphi(\psi) = \text{argmin}_{f \in \mathcal{E}_\varphi}\{-2\langle \psi,\ f\rangle + ||f||_2^2\}\end{aligned} \tag{4}$$

that takes any measure ψ and projects it onto $\mathcal{E}_\varphi$. This operator is motivated by L_2 projection: Whenever $\psi \in L_2(\mathbb{R})$, our operator is equivalent to $P_\varphi(\psi) = \text{argmin}_{f \in \mathcal{E}_\varphi}\{||\psi - f||_2\}$. Here $\langle \psi,\ f\rangle$ denotes the inner product $\int_{\mathbb{R}} \psi(x)\bar{f}(x)\ dx$, and $||f||_2^2 = \langle f,\ f\rangle$. When ψ is a measure, we interpret the inner product as $\langle \psi,\ f\rangle = \mathbb{E}_\psi\left[\bar{f}\right]$. Informally, $P_\varphi(\psi)$ exists and is uniquely defined because it is a projection onto a closed convex subset of the Hilbert space $L_2(\mathbb{R})$. A more precise existence proof is given in the appendix (Proposition A.1).

Under this notation, our estimator is

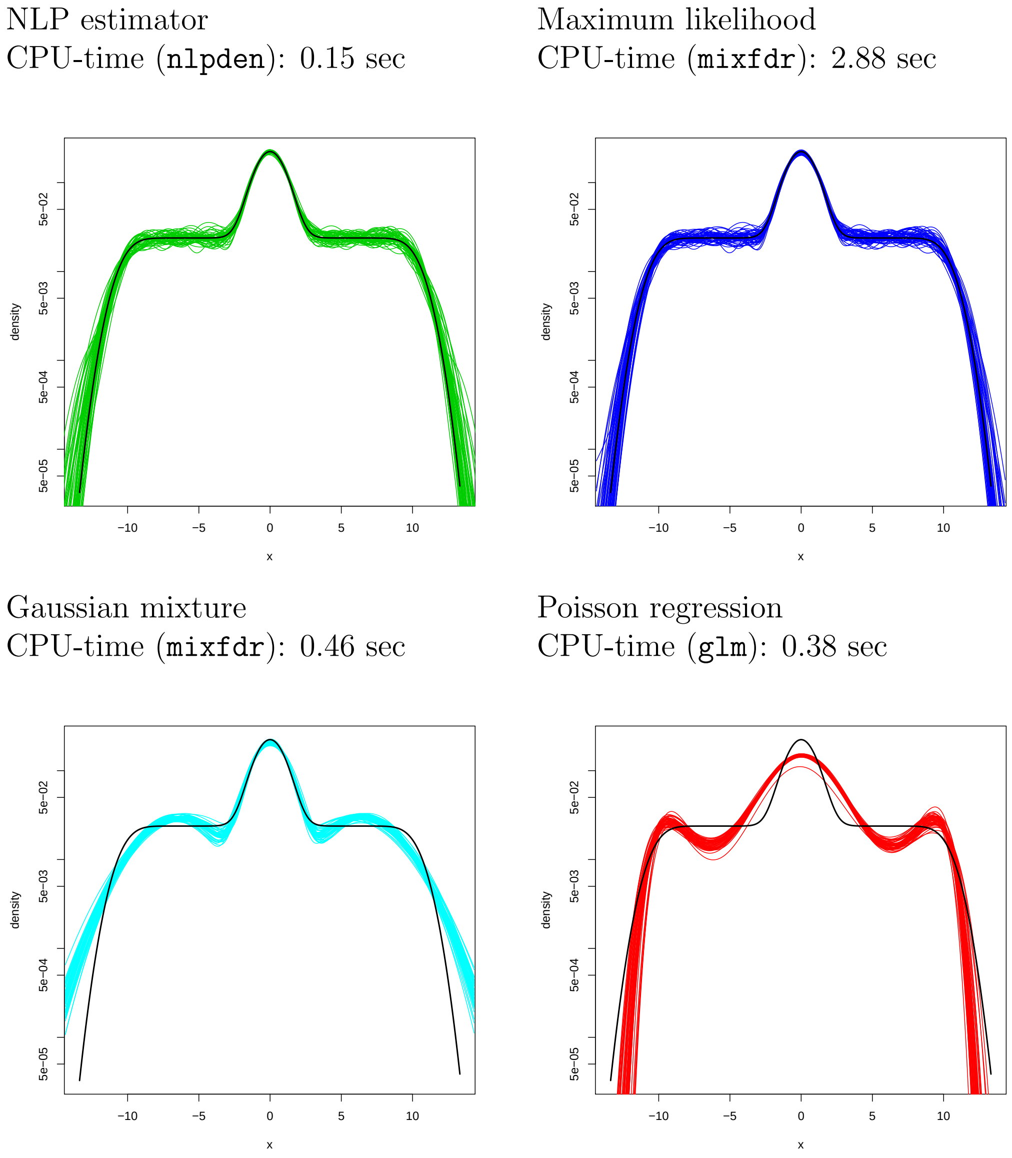
$$\hat{f}_{NLP} = P_\varphi\left(\frac{1}{n}\sum_{i=1}^{n} 1(\{\ .\ = X_i\})\right), \tag{5}$$

i.e. the L_2 projection of the empirical distribution onto the space $\mathcal{E}_\varphi$ of acceptable marginal densities. The subscript NLP stands for "non-linear projection"; we use it to emphasize that $\hat{f}$ is obtained by constrained projection onto a non-linear space.

In this paper, we establish the following properties of our estimator $\hat{f}_{NLP}$: (a) The projection operator P_φ is stable and well-behaved over $\mathcal{D}(\mathbb{R})$, and P_φ is uniformly continuous under smoothing operations. (b) The estimator $\hat{f}_{NLP}$ is consistent. Moreover, under moment conditions on f, we show that it attains a quasi-parametric convergence rate under L_2 norm. (c) The procedure is robust to model mis-specification. Even if f is not actually in the set $\mathcal{E}_\varphi$, $\hat{f}_{NLP}$ will converge—under a global loss function—to the best possible estimate $\bar{f} = P_\varphi f$ at the same rate as when the model was correctly specified. (d) The optimization problem (4) can easily be solved numerically. In fact, the problem can be written as a fairly low-dimensional quadratic program, and so can be solved efficiently using off-the-shelf software. (e) The algorithm for computing $\hat{f}_{NLP}$ does not require an initial guess for $\hat{f}$, and has no sensitive tuning parameters such as bandwidth or degrees of freedom. In general, the advantage of the NLP estimator is that it is almost as accurate as the full maximum likelihood (ML) estimator, but can be computed more efficiently.[1]

In Table 1, we test a suite of estimators in a simulation study where the prior measure g is the mixture of a point mass at zero and a uniform density on $[-10, 10]$; heuristically, this can be thought of as a mixture of 'null' points at zero and 'interesting' points away from zero. We notice that the NLP estimator and the ML estimator have roughly equivalent accuracy, but the former runs almost 20× faster than the latter. Meanwhile, two commonly used alternatives, namely Gaussian mixtures and Poisson regression, advocated by Muralidharan [2010] and Efron [2007] respectively, are noticeably biased for this problem: In particular, the Gaussian mixture systematically overshoots the tails, while Poisson regression undershoots them. We discuss our simulation methodology in more detail in section 4. However, this example shows that our estimator can achieve high accuracy at low computational cost.

[1] As remarked by Omkar Muralidharan, the NLP estimator can also be used in conjunction with the maximum likelihood estimator. One of the biggest difficulties with non-parametric maximum likelihood estimation is that the algorithms used to compute it are sensitive to initialization. Thus, we could use the NLP estimator as a quick and robust way to get a good initialization for maximum likelihood estimation.



NLP estimator
CPU-time (`nlpden`): 0.15 sec

Maximum likelihood
CPU-time (`mixfdr`): 2.88 sec

Gaussian mixture
CPU-time (`mixfdr`): 0.46 sec

Poisson regression
CPU-time (`glm`): 0.38 sec

Table 1: Comparison of four density estimators, run on $N = 2'000$ data points drawn independently from the mixture:

$$X_i \sim \mu_i + \varepsilon_i \text{ with } \mu_i \sim \frac{1}{2}1(\{. = 0\}) + \frac{1}{2}U([-10, 10]) \text{ and } \varepsilon_i \sim \mathcal{N}(0, 1).$$

The plot shows 50 simulation runs for each estimator, as well as the true target density as a thick black line. All estimators were run with default tuning parameters, most importantly $J = 3$ for the Gaussian mixture (J indicates the number of prior components) and $df = 7$ for the Poisson regression. The maximum likelihood estimator is approximate, and was computed as a Gaussian mixture with $J = 20$. CPU-time indicates the average time required to perform one simulation run on the author's laptop.

Finally, our estimator $\hat{f}_{NLP}$ has an explicit representation as $\hat{f} = \varphi * \hat{g}$, and so our procedure implicitly provides an estimate $\hat{g}$ for the distribution of the μ_i. We do not attempt to study the asymptotic properties of $\hat{g}$ explicitly here, as it is known [Carroll and Hall, 1988, Fan, 1991] that minimax convergence rates in the Gaussian deconvolution problem are extremely bad. However, in simulations, $\hat{g}$ appeared to perform just as well or better than kernel based methods for estimating g [e.g. Butucea and Comte, 2009, Comte et al., 2009, Stefanski and Carroll, 1990]. Moreover, unlike kernel estimates for g which may go negative, $\hat{g}$ is guaranteed to correspond to a probability distribution. Thus, studying the behavior of the deconvolution estimate $\hat{g}_{NLP}$ seems like a promising follow-up to the current work.

1.1 Related Methods

Our estimator can be seen as a generalization of minimum distance estimators as pioneered by Wolfowitz [1957], and when restricted to square-integrable densities $\hat{f}_{NLP}$ is in fact the minimum distance density estimator under L_2 norm. Beran and Millar [1994] used minimum distance methods to estimate the distribution of random coefficients in a regression model; these methods have also been used by e.g. Cutler and Cordero-Brana [1996] and Titterington [1983] to fit finite mixture models. Typically, the main motivation for studying minimum distance estimators is that they are robust: In the parametric case, Donoho and Liu [1988] show that these estimators have optimal robustness properties under general conditions. While we do establish robustness properties for $\hat{f}_{NLP}$ (see section 2.3), our main motivation for introducing $\hat{f}_{NLP}$ is its computational tractability.

The least-squares penalty used by our estimator is related to smoothing splines. Both estimators attempt to fit a constrained smooth curve to the empirical distribution under L_2 penalty, their difference being the nature of the constraint imposed on the fitted density. Smoothing splines penalize the curvature of the fitted density, while the NLP estimator replaces this curvature penalty with a shape constraint that comes directly from the Gaussian assumption:

$$\hat{f}_{SPLINE} = \text{argmin}_f \left\{ -\frac{2}{n}\sum_i f(X_i) + ||f||_2^2, \text{ subject to } \int_{\mathbb{R}} \left(f''(x)\right)^2 \, dx < C \right\}, \text{ vs.} \tag{6}$$

$$\hat{f}_{NLP} = \text{argmin}_f \left\{ -\frac{2}{n}\sum_i f(X_i) + ||f||_2^2, \text{ subject to } f = \varphi * g \right\}.$$

In other words, our estimator can be seen as an offshoot of smoothing splines tailored specifically to the Gaussian noise setting.

More generally, our method fits into the class of shape-constrained density estimators. In other contexts, it can be useful to estimate a density f under the constraint that $\hat{f}$ be for example monotone [Durot et al., 2012, Grenander, 1956], convex [Groeneboom et al., 2001], or log-concave [Cule et al., 2010, Dümbgen et al., 2011, Walther, 2009].

2 Theoretical Results

In this section, we outline our main theoretical results concerning $\hat{f}_{NLP}$. For simplicity, we only focus on the Gaussian noise case $\varepsilon_i \sim \mathcal{N}(0,1)$. However, our results hold for quite general additive noise $\varepsilon_i \sim H$, where H is a distribution with density h. The consistency result only requires that $h \in L_2(\mathbb{R})$, while the rate of convergence result requires that h and all its derivatives be individually bounded. All the proofs are given in the appendix. For notational simplicity, we will drop the subscript φ from P_φ and $\mathcal{E}_\varphi$ from now on.

2.1 Consistency and Uniform Continuity

The form of the projection operator P from (4) does not necessarily inspire confidence in the stability of P over general probability distributions. After all, the definition of P in $\mathcal{D}(\mathbb{R})$ is merely a formal extension of a definition that was meaningful and motivated in L_2. Because $\mathcal{E}$ is a convex set, our projection operator P shrinks elements of L_2 towards each other (under the L_2 norm) and so P must clearly be well-behaved over L_2. However, it is not immediately clear that P is well behaved over probability distributions without a density.

The following result aims to dispel any such concerns, as it shows that P is uniformly continuous under small-scale smoothing for all probability measures $\psi \in \mathcal{D}(\mathbb{R})$. A simple and useful corollary of the result is that if we approximate any probability measure ψ with a histogram $H_w(\psi)$, then as the bin width w goes to zero, the projection operator $PH_w(\psi)$ converges uniformly to $P\psi$, no matter how spiky ψ may be. In our algorithm, we exploit this fact by representing the empirical distribution as a narrow bin-width histogram, which is extremely convenient from a computational point of view.

Theorem 2.1 (Stability with Respect to Smoothing). *Let K be a smoothing kernel with a finite second moment and $K_h(x) = h^{-1}K(x/h)$ for any bandwidth $h > 0$. Let P be the projection operator defined in* (4)*. Then,*

$$\lim_{h\to 0} \sup_{\psi \in \mathcal{D}(\mathbb{R})} ||P(K_h * \psi) - P(\psi)||_2 = 0,$$

that is, given a small enough h, we can approximate $P\psi$ with $P(K_h * \psi)$ uniformly in ψ.

One consequence of this theorem is that $\hat{f}_{NLP}$ is consistent. This follows directly from the fact that kernel smoothers with appropriately decaying bandwidths are consistent for functions with bounded derivatives.

Corollary 2.2 (Consistency). *Let X_1, ..., X_n be independently sampled from a distribution with density $f_0 \in \mathcal{E}$, and let $\hat{f}_{NLP}$ be the estimator from* (5)*. Then,*

$$\lim_{n\to\infty} ||\hat{f}_{NLP} - f_0||_2 =_p 0,$$

meaning that $\hat{f}_{NLP}$ converges to f_0 in probability under the L_2 norm.

We can also use the theorem to establish a more abstract uniform continuity result. A popular measure of closeness between probability measures is the Mallows distance (also known as the Wasserstein or earth mover's distance). Suppose that F_X and F_Y are in $\mathcal{P}_1$, the space of probability distributions with finite expectation. Then, the Mallows distance between F_X and F_Y is

$$d(F_X, F_Y) = \inf_{(X,Y)\sim G} \mathbb{E}_G \left[|X - Y|\right] \tag{7}$$

where G ranges over all probability distributions on $\mathbb{R}^2$ with marginals F_X and F_Y; Bickel and Freedman [1981] show that d is a metric over $\mathcal{P}_1$. As an outgrowth of Theorem 2.1, we can show that our projection map P is uniformly continuous under the Mallows metric. Samworth and Yuan [2012] prove a related result for their ICA projection.

Corollary 2.3 (Uniform Continuity). *The restriction $P : (\mathcal{P}_1,\, d) \to (L_2,\, \|.\|_2)$ of the projection map P defined in* (4) *is uniformly continuous.*

2.2 Rate of convergence

We can obtain rate of convergence results for $\hat{f}_{NLP}$ using general results from M-estimation theory. In the language of M-estimation, we can write

$$\hat{f}_{NLP} = \text{argmin}_{f\in\mathcal{E}} \left\{ \frac{1}{n} \sum_{i=1}^{n} m_f(X_i) \right\},$$

where $m_f(\psi) = -2\langle \psi, f\rangle + ||f||_2^2$. We interpret $m_f(X_i)$ as $-2f(X_i) + ||f||_2^2$.

The key technical step in bounding rates of convergence for M-estimators is to derive a maximal inequality for the empirical process

$$\mathbb{G}_n(m_f) = \sqrt{n} \left(\frac{1}{n} \sum_{i=1}^{n} m_f(X_i) - \mathbb{E}[m_f(X_i)] \right),$$

which measures how far the function m_f diverges from its mean. We give such a maximal inequality below. Because the supremum of $\mathbb{G}_n$ is not necessarily measurable, we need to formulate the result in terms of outer expectation $\mathbb{E}^*$. The outer expectation of a function f is the infimum of $\mathbb{E}\left[\tilde{f}\right]$ over all measurable functions $\tilde{f}$ that dominate f and have a well-defined expectation.

Lemma 2.4 (Maximal Inequality). *Let f_0 be a density in $\mathcal{E}$, and let $\{X_i\}$ be drawn independently from the distribution F_0 with density f_0. Then, for any $k \in \mathbb{N}^*$,*

$$\mathbb{E}^*_{F_0} \left[\sup_{\{f\in\mathcal{E}: ||f-f_0||_2<\delta\}} \left|\mathbb{G}_n\left(m_f - m_{f_0}\right)\right| \right] = \mathcal{O}\left(\delta^{\frac{2k-1}{2k}} + \frac{\delta^{-1/k}}{\sqrt{n}} \right),$$

provided that f_0 has rapidly decaying tails, i.e. the tails of f_0 decay faster rate than $|x|^{-a}$ for all $a > 0$. Here, $\mathbb{E}^*_{F_0}$ denotes outer expectation with respect to F_0, and $a(n) = \mathcal{O}(b(n))$ means that, for some $C \geq 0$, $|a(n)| \leq C\,|b(n)|$ for all n.

This maximal inequality can then be transformed into a rate of convergence result. For technical reasons, our proof only works when f_0 has rapidly decaying tails, which is equivalent to F_0 having finite moments of all orders. This should not be a big problem in practice, however, since empirical Bayes methods are usually applied to distributions that are only slightly more dispersed than the normal $\mathcal{N}(0,1)$ distribution.

Theorem 2.5 (Rate of Convergence)**.** *Let $f_0 \in \mathcal{E}$ have rapidly decaying tails as in Lemma 2.4, and let $\hat{f}_{NLP}$ be the estimator* (5)*. Then, for any $\alpha > 0$,*

$$\lim_{n\to\infty} n^{\frac{1}{2(1+\alpha)}} \cdot ||\hat{f}_{NLP} - f_0||_2 =_p 0.$$

Thus, $\hat{f}_{NLP}$ converges to f_0 faster rate than $(1/\sqrt{n})^{1-\varepsilon}$ for all $\varepsilon > 0$; in other words, $\hat{f}_{NLP}$ can get arbitrarily close to the 'parametric' convergence rate $1/\sqrt{n}$ although it may never reach that rate. This kind of convergence rate is often called 'quasi-parametric', and is typical in smooth functional estimation problems [Ibragimov, 2001, gives examples of exact minimax rates under the assumption that f admits an analytic continuation]. The minimax rate for nonparametric estimation of the density of a Gaussian mixture is [Kim, 2012]

$$\inf_{\hat{f}_n} \sup_{f\in\mathcal{E}} \mathbb{E}_{n,f} \left\| \hat{f}_n - f \right\|_2^2 \asymp \frac{\sqrt{\log n}}{n}; \tag{8}$$

the notation means that the ratio of the left- and right-side expressions is bounded above and below by non-zero constants. This rate of convergence is attained by properly chosen kernel estimators [Zhang, 1997]. The standard proof techniques used to establish Theorem 2.5 were not tight enough establish that $\hat{f}_{NLP}$ attains this minimax rate. It would be an interesting topic for further research to see whether $\hat{f}_{NLP}$ in fact attains it.

2.3 Robustness under model mis-specification

Empirical Bayes techniques are often used in cases where the Gaussian noise model is known to hold approximately, but not exactly. For example, it is common to take the X_i to be z-values from a two-sample test, and to interpret μ_i as a measure of effect size. In this case, the Gaussian approximation (1) is very accurate for small effects, but deteriorates somewhat as effect sizes get large.

Thus, it is important to show that any estimator $\hat{f}$ used to estimate the density of the X_i is robust to modest model mis-specification. The following theorem provides such a guarantee. If the model is mis-specified and f_0 is not actually in $\mathcal{E}$, then all our previous results about the convergence of $\hat{f}_{NLP}$ hold provided we replace the target f_0 with the best possible estimate within $\mathcal{E}$, namely $\bar{f}_0 = Pf_0$. Note that the rate of convergence result is stated in terms of a global measure of loss and so does not necessarily tell us how $\hat{f}_{NLP}$ reacts locally to small-scale model mis-specification (see Jankowski and Wellner [2012] for a discussion of this point in the context of the Grenander estimator).

Theorem 2.6 (Robustness)**.** *Let X_0, ..., X_n be independently sampled from a distribution F_0 with finite expectation and a bounded density f_0, and let $\bar{f}_0 = Pf_0$. Then, the projection estimator $\hat{f}_{NLP}$ is consistent in the sense that*

$$\lim_{n\to\infty} ||\hat{f}_{NLP} - \bar{f}_0||_2 =_p 0.$$

Moreover, if f_0 has rapidly decaying tails, then

$$\lim_{n\to\infty} n^{\frac{1}{2(1+\alpha)}} ||\hat{f}_{NLP} - \bar{f}_0||_2 =_p 0$$

for all $\alpha > 0$.

3 Computation

Our estimator $\hat{f}_{NLP}$ is obtained by projecting the empirical distribution onto the space of possible densities in a Gaussian noise model. Specifically, if we write $\delta_n(X)$ for the empirical distribution,

$$\hat{f}_{NLP} = \text{argmin}_{f\in\mathcal{E}}\{||f - \delta_n(X)||_2\}, \tag{9}$$

where $\mathcal{E}$ from (3) is (the closure under L_2 norm of) the set of densities that can be written as $\varphi * g$ for some probability measure g. Of course, $\delta_n(X)$ is not actually in L_2, and we addressed this problem formally in our definition (4) of the projection operator P. However, as shown in Theorem 2.1, $\hat{f}$ is stable with respect to small-scale smoothing. Thus, in practice, we can work with a smoothed version of $\delta_n(X)$, and avoid integrability issues. Our algorithm represents $\delta_n(X)$ with a narrow bin-width histogram.

We have emphasized that an advantage of our estimator $\hat{f}_{NLP}$ is that it can be written as a quadratic program. A quadratic program is a minimization problem of the form

$$\begin{aligned} &\text{Minimize: } \frac{1}{2}x^TQx + c^Tx \\ &\text{Subject to: } Ax \leq b \text{ and } Ex = d, \end{aligned}$$

where Q is a positive definite matrix, A and E are linear transformations, and b, c, and d are column vectors; the minimization is over x. The advantage of formulating a problem as a quadratic program is that such programs can be solved efficiently using off-the-shelf software, such as the `solve.QP` program provided by the `quadprog` library for the `R` programming language.

Because convolution is a linear operation, we confirm that $\hat{f}_{NLP}$ is in fact a quadratic program by writing it as $\hat{f}_{NLP} = \varphi * \hat{g}_{NLP}$, where $\hat{g}_{NLP}$ is the solution to

$$\begin{aligned} &\text{Minimize: } -2\langle\delta_n(X),\ \varphi * g\rangle + ||\varphi * g||_2^2 \\ &\text{Subject to: } g(\mu) \geq 0 \text{ for all } \mu, \text{ and } \int g(\mu)\ d\mu = 1. \end{aligned}$$

In order to make the quadratic program formulation complete, we can think of g as a discrete function taking values over a fine grid; in this case, g can be represented as a vector in $\mathbb{R}^M$ where M is the number of grid points. In the theoretical section we allowed for $\int_{\mathbb{R}} g(x)\ dx \leq 1$ to ensure that $\mathcal{E}$ is closed under L_2 norm; in practice, however, we only work on finite intervals, and so we can use $\int g = 1$ as our constraint because we do not need to worry about probability mass escaping to infinity.

Now, although this expression is a quadratic program, it is still not ideal from a computational point of view. The difficulty is that its dimension scales with the number M of grid points used to describe g, and so it can get computationally intractable as the number of grid points gets large. Thankfully, we can avoid this problem by moving into Fourier space: In our problem, most of the interesting signal is concentrated in the low frequencies, while the high frequencies are submerged by noise. Once we switch to Fourier space, we can throw out most of these high frequencies without losing essentially any information. This leads to a substantial dimensionality reduction, and makes our algorithm fast.

If we assume that $\delta_n(X)$ is in L_2, we can transform the problem into Fourier space using Plancherel's identity

$$||\mathcal{F}f - \mathcal{F}\delta_n(X)||_2 = ||f - \delta_n(X)||_2,$$

where $\mathcal{F}$ stands for the Fourier transform. Our optimization problem then becomes

$$\mathcal{F}\hat{f}_{NLP} = \text{argmin}_{\zeta \in L_2(\mathbb{R})} \left\{ ||\zeta - \mathcal{F}\delta_n(X)||_2 : \mathcal{F}^{-1}\left(\frac{\zeta}{\varphi}\right) \in \mathcal{D}(\mathbb{R}) \right\}, \tag{10}$$

where $\mathcal{D}(\mathbb{R})$ from (2) is (a closure of) the space of all probability measures on $\mathbb{R}$. Here, we took advantage of the fact that convolution becomes multiplication in Fourier space and that φ is its own Fourier transform; thus $f = \varphi * g$ if and only if $\mathcal{F}f = \varphi \cdot \mathcal{F}g$. We can show that after transformation into Fourier space our problem is still a quadratic program, but now with a manageable number of dimensions. We provide an R implementation of the NLP estimator in the package nlpden. R-style pseudo-code for our algorithm is given in Procedure 1.

Procedure 1 Computes the projection estimate $\hat{f}_{NLP}$. FFT stands for Fast Fourier Transform, while the main minimization step is solved by quadratic programming. Multiplication and division are computed component-wise.

```
# params
#   data: the raw observations

data.histogram <- make_histogram(data)
data.fft <- fft(data.histogram)
deconvolution.coeffs <- sqrt(2 * PI) * exp(x^2 / 2)
estimate.fft <-
  MINIMIZE(dummy_var):
    squared_distance(data.fft, dummy_var)
  SUBJECT TO:
    is_non_negative
      inverse_fft(deconvolution.coeffs * dummy_var)
    AND is_one
      sum(inverse_fft(deconvolution.coeffs * dummy_var))
estimate.histogram <- inverse_fft(estimate.fft)
```

Finally, we emphasize that our algorithm is stable in the limit where the bin-width of the data histogram goes to zero. This is a direct consequence of Theorem 2.1, which guarantees that our procedure is uniformly stable under small smoothing operations. The `nlpden` package uses, by default, a bin width of $0.01\,\sigma$ where σ is the standard-deviation of the noise term; the only reason we don't use an even smaller bin width is computational efficiency.

4 Simulation Study

To test our NLP estimator in practice, we matched it up against three commonly used density estimators: Poisson regression, Gaussian mixtures, and maximum likelihood.

Poisson regression is advocated by Efron [2007] as a general-purpose density estimator for empirical Bayes analysis. The idea of Poisson regression for density estimation [e.g., Efron and Tibshirani, 1996] is to model the density f as a natural log-spline, or, more simply, as

$$f(x) = \exp\left[\sum_{k=0}^{K} \alpha_k x^k\right].$$

The latter density estimation problem then reduces to a generalized linear model, and we can obtain the maximum likelihood estimate for f using standard software such as `glm`. Following Efron [2007] and Efron's `R` package `locfdr`, we set the tuning parameter to $K = 6$, which is equivalent to setting the degrees of freedom to $df = 7$.

Muralidharan [2010] showed that finite Gaussian mixtures can be very useful for density estimation even in an infinite mixture setup. We assume

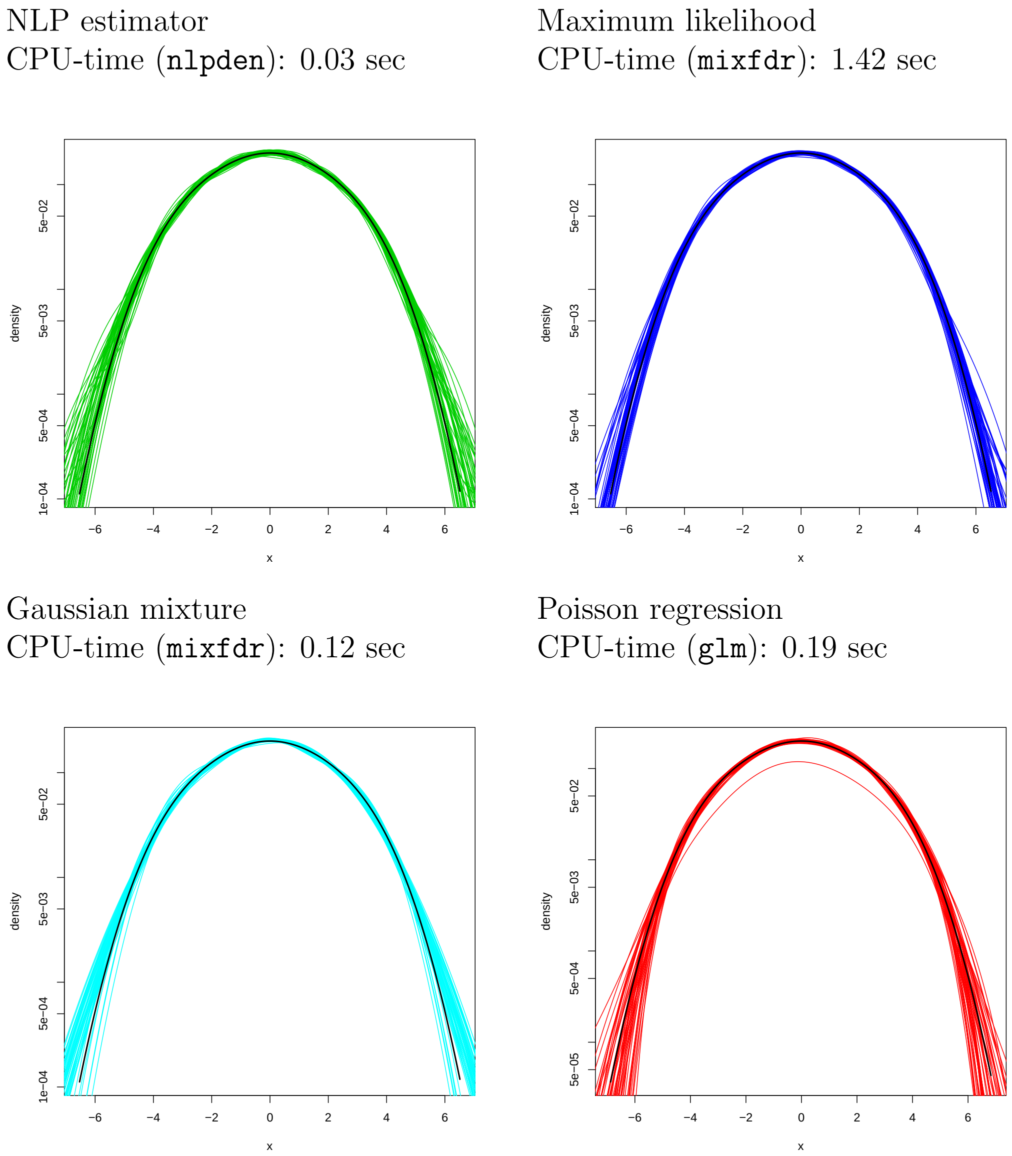
$$f(x) = \sum_{j=1}^{J} p_j \varphi_{\sigma_j}(x - a_j), \text{ where } \sum_j p_j = 1 \text{ and } \sigma_j \geq 1,$$

and φ_{σ_j} is the centered normal density with variance σ_j^2. We fit the mixture using Muralidharan's `R` package `mixfdr`, which implements the EM-algorithm starting from 5 different automatically generated initialization states. Following Muralidharan [2010], we use $J = 3$ components in our mixture.

Maximum likelihood (ML) methods for density estimation have been recently advocated by Jiang and Zhang [2009] and Zhang [2009]. The ML estimator is probably the most accurate estimator available for density estimation under Gaussian noise. The main drawback of ML estimation is that, for this problem, it is extremely demanding computationally. Jiang and Zhang [2009] run an EM-algorithm over a fine grid to compute the ML-estimator. In the hope of getting a computational speed-up, we took a slightly less general approach and approximated the ML-estimator as a Gaussian mixture with 20 components using the `mixfdr` program.

The goal of our simulation study was to evaluate overall performance of the density estimators, rather than performance along some tightly constrained metric such as L_2 error or likelihood. For this reason, we present simulation results in a graphical form, and plot the fitted densities for each simulation run. For empirical Bayes applications, it is important to fit the tails of the density well; we display the y axis on a log scale so that we can inspect tail fit more carefully.[2]

[2] For reported run times, our goal was to give reasonable estimates for run times in 'regular' use, rather



NLP estimator
CPU-time (`nlpden`): 0.03 sec

Maximum likelihood
CPU-time (`mixfdr`): 1.42 sec

Gaussian mixture
CPU-time (`mixfdr`): 0.12 sec

Poisson regression
CPU-time (`glm`): 0.19 sec

Table 2: Comparison of four density estimators, run on $N = 2'000$ data points drawn independently from the distribution:

$$X_i \sim \mu_i + \varepsilon_i \text{ with } \mu_i \sim G \text{ and } \varepsilon_i \sim \mathcal{N}(0,1),$$

$$\text{such that } g(\mu) = \frac{(4 - |\mu|)_+}{16}.$$

The plot shows 50 simulation runs for each estimator, as well as the true target density as a thick black line. All estimators were run with default tuning parameters, most importantly $J = 3$ for the Gaussian mixture (J indicates the number of prior components) and $df = 7$ for the Poisson regression. The maximum likelihood estimator is approximate, and was computed as a Gaussian mixture with $J = 20$. CPU-time indicates the average time required to perform one simulation run on the author's laptop.

Our simulation distributions were motivated by the local false discovery rate problem [Efron et al., 2001]. In the context of the Gaussian noise model, the goal of false discovery analysis is to discern null effects with $\mu_i \approx 0$ from "interesting" effects with $|\mu_i| \gg 0$. Typically, empirical Bayes analysis is much easier to perform when most μ_i are exactly zero and the others are far from zero, i.e. when the vector of the μ_i is sparse. In our simulation study, we first tried a distribution where this sparsity assumption in fact holds, and then tried a second one with no sparsity at all: For our first simulation (Table 1), we let the prior g be the mixture of a point mass at zero and a uniform distribution on $[-10, 10]$, while for the second (Table 2), g had a triangle-shaped density with support on $[-4, 4]$.

In both examples (Tables 1 and 2), we see that our NLP estimator performs roughly as well as maximum likelihood, but at a 20–50× reduction in CPU-time. Poisson regression does quite well for the triangle prior, but does not do well at all for the first mixture. In particular, it badly undershoots the tails. We tried increasing the degrees of freedom for the Poisson regression, but this didn't help much; rather, the fit started to become very unstable after we introduced more than 15 degrees of freedom. The 3-component Gaussian mixture somewhat overshoots the tails in both cases, the effect being more pronounced in Table 1 than in Table 2.

Acknowledgment

I am grateful for helpful discussions with Brad Efron, Omkar Muralidharan, Amir Najmi, Suzanne de Treville and Guenther Walther. I would also like to thank two anonymous reviewers for their useful comments and for a conjecture that led to Corollary 2.3. This work was supported by a B.C. and E.J. Eaves Stanford Graduate Fellowship.

A Proofs

To make sure that convolution is well specified whenever ψ is a measure, we define convolution as $(K * \psi)(x) = \mathbb{E}_\psi[K(x - \mu)]$, where μ is distributed according to ψ.

Proposition A.1. *The projection operator P as described in* (4) *is well-defined and unique.*

Proof. Let $\psi \in \mathcal{D}(\mathbb{R})$, and let $\Delta_\psi : \mathcal{E} \to \mathbb{R}$ with $\Delta_\psi(f) = -2\langle \psi, f \rangle + \|f\|_2^2$ be the objective function from (4). We need to show that Δ_ψ attains its minimum value at a unique $f^* \in \mathcal{E}$. Let $L = \inf_{f \in \mathcal{E}} \{\Delta_\psi(f)\}$. Because Δ_ψ is strictly convex and bounded below on $\mathcal{E}$, we see that L is finite and that Δ_ψ can attain its infimum L at most once. It remains to show that there exists a solution $f^* \in \mathcal{E}$ satisfying $\Delta_\psi(f^*) = L$.

than to give maximally optimized run times for each estimator. Here, each estimator could easily be made substantially faster by cutting some corners. The `mixfdr` package always uses 5 starts to estimate maximum likelihood and Gaussian mixtures, and so we could cut the run time by 5 by only selecting 1 start and hoping it works. Similarly, both `nlpden` and our implementation of Poisson regression use a histogram with bin width 0.01 to compute estimates. Arguably, this bin width is needlessly small, and we could speed up the algorithm a lot (asymptotically by a factor 10 for `nlpden`) by moving to a bin width of 0.1. However, instead of tuning each estimator for the simulation study, we chose to use each estimator with 'default' computational parameters.

The following argument is closely adapted from the proof of Hilbert's projection theorem in Rudin [1987]. Let $\{f_n\}_{n=1,\,2,\,\ldots}$ be a sequence of functions satisfying $\Delta_\psi(f_n) < L + 1/n$ for each integer n. Then, for any m and n,

$$\begin{aligned}\frac{1}{2}\|f_m - f_n\|_2^2 &= \|f_m\|_2^2 + \|f_n\|_2^2 - 2\left\|\frac{f_m + f_n}{2}\right\|_2^2 \\ &= \Delta_\psi(f_m) + \Delta_\psi(f_n) - 2\,\Delta_\psi\left(\frac{f_m + f_n}{2}\right) \\ &< \frac{1}{m} + \frac{1}{n},\end{aligned}$$

where on the last line we used the fact that $\mathcal{E}$ is convex and so $(f_m + f_n)/2 \in \mathcal{E}$. Thus, $\{f_n\}$ is a Cauchy sequence. Because $L_2(\mathbb{R})$ is complete and $\mathcal{E}$ is closed under L_2 norm, $\{f_n\}$ has a limit $f^* \in \mathcal{E}$. By continuity of Δ_ψ we conclude that $\Delta_\psi(f^*) = L$. □

Proof of Theorem 2.1. Let $\chi(\psi)(t) = \mathbb{E}_\psi\left[e^{itX}\right]$ be the characteristic function of ψ, and, for any $c > 0$, let T_c be the spectral truncation operator

$$T_c(\psi) = \chi^{-1}\left[1([-c, c]) \cdot \chi(\psi)\right],$$

where 1 is an indicator function.

With this notation, we can bound our expression of interest with

$$\begin{aligned}&||P(K_h * \psi) - P(\psi)||_2^2 \\ &\quad\leq 3 \cdot ||(P \circ T_c)(K_h * \psi) - (P \circ T_c)(\psi)||_2^2 \\ &\quad+ 3 \cdot ||P(K_h * \psi) - (P \circ T_c)(K_h * \psi)||_2^2 \\ &\quad+ 3 \cdot ||P(\psi) - (P \circ T_c)(\psi)||_2^2.\end{aligned} \tag{11}$$

We start by giving a uniform bound for the last summand, which also applies to the second summand. This bound exploits the convexity of the set $\mathcal{E}$.

It is well known that $\chi(\varphi * \psi) = \chi(\varphi)\chi(\psi)$, that $\chi(\varphi)(t) = e^{-t^2/2}$, and that $||\chi(\psi)||_\infty \leq ||\psi||_1$. Given these observations, we can use Parseval's theorem to show that for any $c > 0$ and $\psi_1,\ \psi_2 \in \mathcal{D}(\mathbb{R})$,

$$\begin{aligned}|\langle P\psi_1,\ \psi_2\rangle - \langle P\psi_1,\ T_c\psi_2\rangle| &= \frac{1}{2\pi}\,|\langle \chi(P\psi_1),\ \chi(\psi_2 - T_c\psi_2)\rangle| \\ &\leq \frac{1}{2\pi}\left|\left\langle e^{-t^2/2},\ 1(\{|t| > c\})\right\rangle\right| \\ &= \sqrt{\frac{2}{\pi}}\Phi(-c).\end{aligned}$$

Thus, for any $\varepsilon > 0$, we can pick a $c > 0$ such that

$$|\langle P\psi_1,\ \psi_2\rangle - \langle P\psi_1,\ T_c\psi_2\rangle| < \varepsilon$$

for all ψ_1 and $\psi_2 \in \mathcal{D}(\mathbb{R})$. Now, with this value of c, we can show that for any $\psi \in \mathcal{D}(\mathbb{R})$,

$$\begin{aligned}
&-2\langle P(T_c\psi),\ \psi\rangle + ||P(T_c\psi)||_2^2 \\
&\quad \leq -2\langle P(T_c\psi),\ T_c\psi\rangle + ||P(T_c\psi)||_2^2 + \varepsilon \\
&\quad \leq -2\langle P\psi,\ T_c\psi\rangle + ||P\psi||_2^2 + \varepsilon \\
&\quad \leq -2\langle P\psi,\ \psi\rangle + ||P\psi||_2^2 + 2\varepsilon.
\end{aligned}$$

Here, the first an the last inequalities were due to our choice of c, while the middle inequality is true since, by the definition of P, we know that

$$P(T_c\psi) = \text{argmin}_{\tilde{\psi}\in\mathcal{E}} -2\left\langle \tilde{\psi},\ P(T_c\psi)\right\rangle + ||\tilde{\psi}||_2^2.$$

If we expand the square

$$||P(T_c\psi)||_2^2 = ||P\psi||^2 + 2\langle P\psi, P(T_c\psi) - P\psi\rangle + ||P(T_c\psi) - P\psi||_2^2,$$

we can write the above inequality as

$$\frac{1}{2}||P(T_c\psi) - P\psi||_2^2 \leq \langle P(T_c\psi) - P\psi,\ \psi - P\psi\rangle + \varepsilon.$$

Finally, since $\mathcal{E}$ is a convex set, we must have

$$\langle P(T_c\psi) - P\psi,\ \psi - P\psi\rangle \leq 0,$$

as otherwise, by convexity, there would be a point on the line connecting $P\psi$ and $P(T_c\psi)$ that is strictly closer to ψ than $P\psi$ under L_2 norm, which would contradict the fact that $P\psi$ is the L_2 projection of ψ onto $\mathcal{E}$. Thus, by picking a large enough c, we can make the last two summands in (11) smaller than ε uniformly over $\mathcal{D}$.

We now move to the first summand. Because the smoothing kernel K has a finite second moment, χK_h converges to 1 as h converges to 0 uniformly on compact intervals of $\mathbb{R}$. For any $\psi \in \mathcal{D}(\mathbb{R})$ we have $||\chi(\psi)||_\infty \leq 1$, and so by Plancherel's theorem,

$$\begin{aligned}
||T_c(K_h * \psi) - T_c\psi||_2^2 &= \frac{1}{2\pi}\int_{-c}^{c} (1 - \chi(K_h)(t))^2 \chi(\psi)(t)^2\ dt \\
&\leq \frac{1}{2\pi}\int_{-c}^{c} ||1 - \chi(K_h)(t)||_2^2\ dt.
\end{aligned}$$

Thus, for any fixed c, we can pick an $h > 0$ that makes this quantity arbitrarily small uniformly over $\psi \in \mathcal{D}(\mathbb{R})$. Finally, P is a projection onto a convex set, and so

$$||P(\psi_1) - P(\psi_2)||_2^2 \leq ||\psi_1 - \psi_2||_2^2$$

for any ψ_1 and ψ_2 in $L_2(\mathbb{R})$. Thus, given any ε, $c > 0$, there is an $h_c > 0$ such that

$$||(P \circ T_c)(K_h * \psi) - (P \circ T_c)(\psi)||_2^2 < \varepsilon$$

for all $0 < h < h_c$ and $\psi \in \mathcal{D}(\mathbb{R})$. □

Proof of Corollary 2.2. For convenience, let

$$\delta_n(X)(x) = \frac{1}{n}\sum_{i=1}^{n} 1(\{x = X_i\})$$

be the empirical distribution. Let $\varphi_h(x) = h^{-1}\,\varphi(x/h)$ be the standard Gaussian kernel with bandwidth h. It is well known [e.g. Rosenblatt, 1971] that, when f_0 and its first two derivatives are bounded (this condition is satisfied here because $f_0 = \varphi * g_0$), there exists a sequence of bandwidths $h_n \to 0$ such that

$$\lim_{n\to\infty} ||\varphi_{h_n} * \delta_n(X) - f_0||_2 =_p 0.$$

Now, since f_0 is in $\mathcal{E}$ and this set is convex, taking the projection $P\hat{f}$ of any estimator $\hat{f}$ for f_0 can only improve the performance of the estimator under L_2 norm, and so

$$||P(\varphi_h * \delta_n(X)) - f_0||_2 \leq ||\varphi_h * \delta_n(X) - f_0||_2. \tag{12}$$

Finally, by Theorem 2.1,

$$\lim_{n\to\infty} ||P(\varphi_{h_n} * \delta_n(X)) - P\delta_n(X)||_2 =_p 0$$

uniformly in X, which implies the desired conclusion. □

Proof of Corollary 2.3. Is is well known [e.g. Bickel and Freedman, 1981] that the Mallows distance as defined in (7) has the following simpler representation:

$$d(F_X, F_Y) = \int_0^1 \left|F_X^{-1}(u) - F_Y^{-1}(u)\right|\, du.$$

Now, let F_X and F_Y probability distributions with finite first absolute moments and corresponding Radon-Nikodym derivatives $\psi_X = dF_X/d\lambda$ and $\psi_Y = dF_Y/d\lambda$ with respect to the Lebesgue measure λ. We need to show that for any $\varepsilon > 0$ there is a $\delta > 0$ such that

$$d^2(F_X, F_Y) < \delta \text{ implies that } \|P(\psi_X) - P(\psi_Y)\|_2^2 < \varepsilon.$$

Let $\varepsilon > 0$ be fixed. By Theorem 2.1, we can pick $h > 0$ such that

$$\|P(\varphi_h * \psi) - P(\psi)\|_2^2 < \varepsilon/4 \text{ for all } \psi \in \mathcal{D}.$$

It follows that

$$\begin{aligned}\|P(\psi_X) - P(\psi_Y)\|_2^2 &\leq \|P(\varphi_h * \psi_X) - P(\varphi_h * \psi_Y)\|_2^2 + \varepsilon/2 \\ &\leq \|\varphi_h * (\psi_X - \psi_Y)\|_2^2 + \varepsilon/2 \\ &= \frac{1}{2\pi}\int_{\mathbb{R}} e^{-h^2t^2}\chi^2(\psi_X - \psi_Y)(t)\; dt + \varepsilon/2.\end{aligned}$$

Here, the second inequality holds because P is a projection onto a convex set, while the last equality is an application of Parseval's theorem. Moreover,

$$\begin{aligned}\chi^2(\psi_X - \psi_Y)(t) &\leq \left(\int_0^1 \left| e^{itF_X^{-1}(u)} - e^{itF_Y^{-1}(u)} \right| \, du \right)^2 \\ &\leq t^2 \left(\int_0^1 \left| F_X^{-1}(u) - F_Y^{-1}(u) \right| \, du \right)^2 \\ &= t^2 \, d^2(F_X, F_Y).\end{aligned}$$

Thus,

$$\|P(\psi_X) - P(\psi_Y)\|_2^2 \leq \frac{d^2(F_X, F_Y)}{4\sqrt{\pi}\, h^3} + \varepsilon/2,$$

which implies the desired result. □

Proof of Lemma 2.4. The main theoretical device used in this proof is *bracketing numbers.* An ε-bracket with respect to the metric $L_2(F_0)$ is a pair of functions $[l, u]$ with $l(x) \leq u(x)$ for all $x \in \mathbb{R}$ such that $\mathbb{E}_{F_0}[(u(x) - l(x))^2] < \varepsilon^2$. The $[l, u]$ bracket contains f if $l \leq f \leq u$. For any set $\mathcal{B}$, we can compute the bracketing number $N_{[\,]}(\varepsilon, \mathcal{B}, L_2(F_0))$, which corresponds to the minimum number of ε-brackets required to cover all the elements of $\mathcal{B}$.

Let $\mathcal{B}_\delta(f_0) = \{\tilde{f} : ||\tilde{f}||_2 < \delta,\ f_0 + \tilde{f} \in \mathcal{E}\}$. Notice that if $\tilde{f} = f - f_0$, then $\mathbb{G}_n(m_f - m_{f_0}) = -2\mathbb{G}_n(\tilde{f})$. Because $f_0 \in \mathcal{E}$ we know that $||f_0||_\infty \leq 1/\sqrt{2\pi}$, and so

$$\mathbb{E}_{F_0}\left[\tilde{f}^2\right] \leq ||f_0||_\infty \cdot ||\tilde{f}||_2^2 < \delta^2. \tag{13}$$

With this bound, we can use Lemma 19.36 of Van der Vaart [2000] to show that

$$\mathbb{E}^*_{F_0}\left[\sup_{\tilde{f} \in \mathcal{B}_\delta(f_0)} \left| \mathbb{G}_n\left(\tilde{f}\right) \right| \right] = \mathcal{O}\left(J_{[\,]} \cdot \left(1 + \frac{J_{[\,]}}{\delta^2 \sqrt{n}} \right) \right),$$

where J is the bracketing integral

$$J_{[\,]} = \int_0^\delta \sqrt{\log N_{[\,]}(\varepsilon, \mathcal{B}_\delta(f_0), L_2(F_0))} \, d\varepsilon.$$

Now, we can write $\tilde{f}$ as $\varphi * (g - g_0)$, where g and g_0 are probability distributions. Thus, for any k, $\tilde{f}$ is k-times differentiable with $\tilde{f}^{(k)} = \varphi^{(k)} * (g - g_0)$, and the k first derivatives of $\tilde{f}$ are bounded by the universal constant $M_k = 2\ \sup\{|\varphi^{(j)}(x)| : x \in \mathbb{R},\ j = 0, ..., k\}$. Thanks to this, we can use Example 19.9 of Van der Vaart [2000] to show that for any $k \in \mathbb{N}^*$,

$$\log N_{[\,]}(\varepsilon, \mathcal{B}_\delta(f_0), L_2(F_0)) \leq M'_k \left(\sum_{z \in \mathbb{Z}} \left(\int_z^{z+1} f_0(x) \, dx \right)^{\frac{1}{1+2k}} dx \right)^{\frac{1+2k}{2k}} \varepsilon^{-\frac{1}{k}},$$

where M'_k is another universal constant and $\mathbb{N}^*$ is the set of positive integers. Because we assumed f_0 to have rapidly decaying tails, the discretized integral over f_0 is finite. Thus, we find that

$$J_{[\,]} = \mathcal{O}\left(\delta^{\frac{2k-1}{2k}}\right),$$

which implies the desired result. □

Proof of Theorem 2.5. Let $k \in \mathbb{N}^*$, $r_n = n^{\frac{1}{2+1/k}}$, and let $S_{j,n}$ be the set of $f \in \mathcal{E}$ with $2^{j-1} < r_n \; ||f - f_0||_2 \leq 2^j$. Now,

$$P^*(r_n \; ||\hat{f} - f_0||_2 > 2^M) \leq \sum_{j \geq M} P^* \left(\inf_{f \in S_{j,n}} \frac{1}{n} \sum_{i=1}^{n} (m_f(X_i) - m_{f_0}(X_i)) \leq 0 \right).$$

We can verify that

$$\mathbb{E}_{F_0}[m_f(X) - m_{f_0}(X)] = ||f - f_0||_2^2, \tag{14}$$

and so

$$\frac{1}{n} \sum_{i=1}^{n} (m_f(X_i) - m_{f_0}(X_i)) = \frac{1}{\sqrt{n}} \mathbb{G}_n(m_f - m_{f_0}) + ||f - f_0||_2^2.$$

By applying Lemma 2.4 with our constant k, we get

$$\begin{aligned}
&P^*(r_n \; ||\hat{f} - f_0||_2 > 2^M) \\
&\quad \leq \sum_{j \geq M} P^* \left(\inf_{f \in S_{j,n}} \mathbb{G}_n(m_f - m_{f_0}) \leq -\sqrt{n} \cdot \inf_{f \in S_{j,n}} ||f - f_0||_2^2 \right) \\
&\quad \leq \sum_{j \geq M} P^* \left(\inf_{f \in S_{j,n}} \mathbb{G}_n(m_f - m_{f_0}) \leq -\sqrt{n} \cdot \frac{4^{j-1}}{r_n^2} \right) \\
&\quad \leq K_k \cdot \frac{r_n^2}{\sqrt{n}} \cdot \left(r_n^{-(1-1/(2k))} + \frac{r_n^{1/k}}{\sqrt{n}} \right) \cdot \sum_{j \geq M} \frac{1}{2^{j-2}} \\
&\quad = \frac{K_k}{2^{M-3}},
\end{aligned}$$

where the second-to-last step is valid by Markov's inequality, and K_k is the constant from Lemma 2.4. Since this bound holds uniformly in n, we can conclude that, for any $k \in \mathbb{N}^*$

$$n^{\frac{1}{2+1/k}} \cdot ||\hat{f}_{NLP} - f_0||_2 = \mathcal{O}_P^*(1),$$

where $\mathcal{O}_P^*(1)$ means 'bounded in probability under outer measure'. Because this result holds for all $k \in \mathbb{N}^*$, we get the desired result. □

Proof of Theorem 2.6. Since our proofs are mainly built on geometric arguments, an extension to mis-specified models is surprisingly straightforward. We begin by establishing consistency. Since F_0 has finite expectation, we know from Bickel and Freedman [1981] that $d\left(\hat{F}_0^{(n)}, F_0\right)$ converges to zero in probability, where $\hat{F}_0^{(n)}$ is the empirical distribution (in fact, almost sure convergence also holds). Thus, by Corollary 2.3, our projection estimator converges in probability to $P(f_0)$.

Now, since f_0 is bounded and so (13) holds up to a constant, we can use the same proof as in Lemma 2.4 to show that for any $k \in \mathbb{N}^*$,

$$\mathbb{E}_{F_0}^* \left[\sup_{\{f \in \mathcal{E}: ||f - \bar{f}_0||_2 < \delta\}} \left| \mathbb{G}_n \left(m_f - m_{\bar{f}_0} \right) \right| \right] = \mathcal{O} \left(\delta^{\frac{2k-1}{2k}} + \frac{\delta^{-1/k}}{\sqrt{n}} \right),$$

provided that f_0 has rapidly decaying tails. With this result, we can mimic the proof of Theorem 2.5 to establish the desired rate of convergence result. The only difference is that we need to replace (14) with

$$\begin{aligned}\mathbb{E}_{F_0}[m_f(X) - m_{\bar{f}_0}(X)] &= ||f - f_0||_2^2 - ||\bar{f}_0 - f_0||_2^2 \\ &= ||f - \bar{f}_0||_2^2 + 2\langle f - \bar{f}_0,\ \bar{f}_0 - f_0\rangle \\ &\geq ||f - \bar{f}_0||_2^2,\end{aligned}$$

where the last statement is true because $\mathcal{E}$ is convex. □

References

R. Beran and P.W. Millar. Minimum distance estimation in random coefficient regression models. *The Annals of Statistics*, pages 1976–1992, 1994.

P.J. Bickel and D.A. Freedman. Some asymptotic theory for the bootstrap. *The Annals of Statistics*, 9(6):1196–1217, 1981.

L.D. Brown and E. Greenshtein. Nonparametric empirical Bayes and compound decision approaches to estimation of a high-dimensional vector of normal means. *The Annals of Statistics*, 37(4):1685–1704, 2009.

C. Butucea and F. Comte. Adaptive estimation of linear functionals in the convolution model and applications. *Bernoulli*, 15(1):69–98, 2009.

R.J. Carroll and P. Hall. Optimal rates of convergence for deconvolving a density. *Journal of the American Statistical Association*, pages 1184–1186, 1988.

F. Comte, Y. Rozenholc, and M.L. Taupin. Penalized contrast estimator for adaptive density deconvolution. *Canadian Journal of Statistics*, 34(3):431–452, 2009.

M. Cule, R. Samworth, and M. Stewart. Maximum likelihood estimation of a multi-dimensional log-concave density. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(5):545–607, 2010.

A. Cutler and O.I. Cordero-Brana. Minimum Hellinger distance estimation for finite mixture models. *Journal of the American Statistical Association*, 91(436):1716–1723, 1996.

D.L. Donoho and R.C. Liu. The "automatic" robustness of minimum distance functionals. *The Annals of Statistics*, 16(2):552–586, 1988.

L. Dümbgen, R. Samworth, and D. Schuhmacher. Approximation by log-concave distributions, with applications to regression. *The Annals of Statistics*, 39(2):702–730, 2011.

C. Durot, V.N. Kulikov, and H.P. Lopuhaä. The limit distribution of the L-infinity error of grenander-type estimators. *The Annals of Statistics*, 40(3):1578–1608, 2012.

B. Efron. Size, power and false discovery rates. *The Annals of Statistics*, 35(4):1351–1377, 2007.

B. Efron. *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Cambridge Univ Pr, 2010.

B. Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011.

B. Efron and R. Tibshirani. Using specially designed exponential families for density estimation. *The Annals of Statistics*, 24(6):2431–2461, 1996.

B. Efron, R. Tibshirani, J.D. Storey, and V. Tusher. Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, 96(456):1151–1160, 2001.

J. Fan. On the optimal rates of convergence for nonparametric deconvolution problems. *The Annals of Statistics*, pages 1257–1272, 1991.

U. Grenander. On the theory of mortality measurement. *Scandinavian Actuarial Journal*, 1956(2):125–153, 1956.

P. Groeneboom, G. Jongbloed, and J.A. Wellner. Estimation of a convex function: Characterizations and asymptotic theory. *The Annals of Statistics*, 29(6):1653–1698, 2001.

I Ibragimov. Estimation of analytic functions. *Lecture Notes-Monograph Series*, pages 359–383, 2001.

H.K. Jankowski and J.A. Wellner. Convergence of linear functionals of the grenander estimator under misspecification. *arXiv preprint arXiv:1207.6614*, 2012.

W. Jiang. On regularized general empirical Bayes estimation of normal means. *Journal of Multivariate Analysis*, 2012.

W. Jiang and C.H. Zhang. General maximum likelihood empirical Bayes estimation of normal means. *The Annals of Statistics*, 37(4):1647–1684, 2009.

I.M. Johnstone and B.W. Silverman. Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. *The Annals of Statistics*, 32(4):1594–1649, 2004.

K.H. Kim. Minimax bounds for estimation of normal mixtures. *arXiv preprint arXiv:1112.4565*, 2012.

R. Koenker and I. Mizera. Convex optimization, shape constraints, compound decisions, and emipircal Bayes rules. Technical report, 2012. URL `http://ysidro.econ.uiuc.edu/~roger/research/ebayes/brown.pdf`.

N. Laird. Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*, 73(364):805–811, 1978.

O. Muralidharan. An empirical Bayes mixture method for effect size and false discovery rate estimation. *The Annals of Applied Statistics*, 4(1):422–438, 2010.

H. Robbins. The empirical Bayes approach to statistical decision problems. *The Annals of Mathematical Statistics*, pages 1–20, 1964.

M. Rosenblatt. Curve estimates. *The Annals of Mathematical Statistics*, 42(6):1815–1842, 1971.

W. Rudin. *Real and complex analysis.* McGraw-Hill, New York, 3rd edition, 1987.

R.J. Samworth and M. Yuan. Independent component analysis via nonparametric maximum likelihood estimation. *arXiv preprint arXiv:1206.0457*, 2012.

L.A. Stefanski and R. J. Carroll. Deconvoluting kernel density estimators. *Statistics: A Journal of Theoretical and Applied Statistics*, 21(2):169–184, 1990.

D.M. Titterington. Minimum distance non-parametric estimation of mixture proportions. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 37–46, 1983.

A. Van der Vaart. *Asymptotic statistics.* Cambridge Univ Pr, 2000.

G. Walther. Inference and modeling with log-concave distributions. *Statistical Science*, 24 (3):319–327, 2009.

J. Wolfowitz. The minimum distance method. *The Annals of Mathematical Statistics*, 28 (1):75–88, 1957.

C.H. Zhang. Empirical Bayes and compound estimation of normal means. *Statistica Sinica*, 7:181–194, 1997.

C.H. Zhang. Generalized maximum likelihood estimation of normal mixture densities. *Statistica Sinica*, 19(3):1297, 2009.