A Statistician's View on Deconvolution and Unfolding

Victor M. Panaretos
École Polytechnique Fédérale de Lausanne, Switzerland

Abstract
We briefly review some of the basic features of unfolding problems from the point of view of the statistician. To illustrate these, we mostly concentrate on the particular instance of unfolding called deconvolution. We discuss the issue of ill-posedness, the bias-variance trade-off, and regularisation tuning, placing emphasis on the important class of kernel density estimators. We also briefly consider basic aspects of the more general unfolding problem and mention some of the points that where raised during the discussion session of the unfolding workshop.

1 Introduction

Unfolding and deconvolution can be seen to arise as variants of the statistical problem of *estimation*, when there is the additional complication of measurement error. In a classical setting, we are able to collect data in the form of realisations of random variables $(X_1, ..., X_n)$. These are often assumed independent and identically distributed according to a cumulative probability distribution F_X belonging to some known class of distributions $\mathcal{F}$. The problem of point estimation can then be formulated as follows:

1. Assume that $\mathcal{F}$ is known but F_X is unknown.
2. Observe a realisation $(x_1, ..., x_n)$ of $(X_1, ..., X_n)$, generated by F_X.
3. Determine which model from $\mathcal{F}$ generated the sample on the basis of $(x_1, ..., x_n)$ (i.e. estimate F_X).

Depending on the degree of specification of the collection $\mathcal{F}$, we may distinguish two broad classes of estimation problems. The first one, called *parametric estimation*, considers collections $\mathcal{F}$ that can be parametrised (put in one-to-one correspondence) by some subset of Euclidean space, $\mathcal{F} = \{F_\theta : \theta \in \Theta \subseteq \mathbb{R}^p\}$. The functional form of the elements of $\mathcal{F}$ is thus completely known, except for a finite-dimensional Euclidean parameter. The second broad class of estimation problems considers collections $\mathcal{F}$ that are only weakly specified, in the sense that they are taken to be subsets of a function space defined through some broad qualitative constraints. For example, $\mathcal{F}$ could be taken to be the collection of distributions possessing densities that are twice continuously differentiable.

An essential difference between the parametric and nonparametric frameworks concerns the effective dimensionality of the problem. In the parametric case, the estimation problem reduces to the determination of a finite dimensional parameter whose dimension remains fixed as the sample size grows. In the nonparametric framework, there is no finite-dimensional reduction of the problem: even if sample size might constrain us to approximate the truth by a function of effectively finite dimension, this dimension will typically increase along with sample size. Note here, that if the dimension of the parameter of a parametric model is allowed to grow with the sample size (i.e. the the more the data, the richer the model we employ), then, as sample size increases, the distinction from a nonparametric model decreases.

Occasionally, the problem –be it parametric, or nonparametric– is further perturbed by measurement limitations. That is, instead of the collection $\boldsymbol{X} = (X_1, ..., X_n)$, we can only observe a proxy $\boldsymbol{Y} = (Y_1, ..., Y_n)$, that results from some random perturbation of $\boldsymbol{X}$, governed by a *measurement error mechanism*:

$$Y_i = g(X_i, \varepsilon_i), \qquad i = 1, ..., n.$$

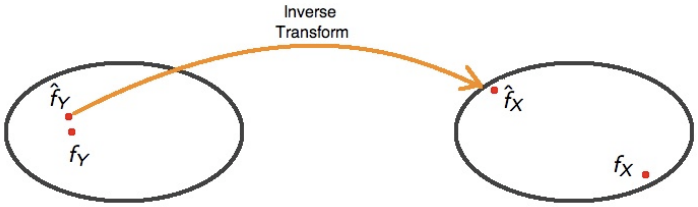

Fig. 1: Schematic illustration of ill-posedness: small errors in the estimation of f_Y may be translated into large errors in estimation of f_X

Here, one typically assumes that the random errors ε_i are mutually independent, identically distributed and independent of the collection $\boldsymbol{X}$, the function g is smooth and that the functions $\{g_t(\cdot) = g(t, \cdot); t \in \mathbb{R}\}$, are invertible with differentiable inverses (i.e. if we knew the input and the response, then we should be able to uniquely and stably determine the error). The error inputs ε_i are unobservable, but their distribution F_ε will be assumed to be completely known. All random variables are assumed real.

It follows that the data we observe are from a distribution F_Y and not the distribution of interest F_X. Assuming that all random variables involved possess density functions denoted by f_X and f_Y, respectively, we may write

$$f_Y(y) = \int f_{Y|X}(y|x) f_X(x) dx = \int h(x, y) f_X(x) dx,$$

where h is connected to the density f_ε of ε and the measurement function g through the change of density formula:

$$h(x, y) = f_{Y|X}(y|x) = \left| J_{g_x^{-1}}(y) \right| \times f_\varepsilon \left(g_x^{-1}(y) \right).$$

Here, J stands for the Jacobian of the transformation. The unfolding problem (or one version of what statisticians call a measurement error model) then consists in estimating the density f_X when one observes realisations from the density f_Y.

In principle, the measurement error problem does not pose real conceptual difficulties in the parametric case because from a qualitative point of view the problem remains the same: the density f_Y will still depend on the original parameter θ, and therefore estimation of θ can be carried out directly in the Y-space without additional complications, at least provided that θ remains identifiable, or the likelihood is not significantly "flattened" (lack, or almost lack, of identifiability can, however, be an issue and would lead to serious complications; one way to address these is by identifiability constraints, which essentially amount to *parametric regularisation*, though we will not pursue this further here).

In the nonparametric case, however, f_X is completely unknown, and hence one cannot escape the measurement error problem and work solely on Y-space. Rather, one will need to first estimate f_Y by some $\hat{f}_Y$, and then attempt to invert the integral transform connecting f_Y with f_Y, using $\hat{f}_Y$ as a proxy for f_Y. This naive approach, however, can lead to serious errors. For well-behaved h, the integral transformation involved will have an discontinuous (unbounded) inverse transform. Consequently, an element of ill-posedness enters the picture, as small errors in the estimation of f_Y may be translated into large errors in estimation of f_X through the inversion process.

If the measurement errors are not independent, then the observed variables $(Y_1, ..., Y_n)$ will no longer constitute an independent random sample, but will instead form a stationary process. Consequently, the integral expression $\int h(x, y) f_X(x) dx$ will still hold marginally for the density of each Y_i, but the joint density will no longer be the product density. Nevertheless, even in such cases, one can attempt to proceed using the same estimators as in the independent case, provided that the dependence among the errors is weak (where the notion of 'weak dependence' can be formalised through appropriate

mixing conditions). The stronger the dependence structure of the errors, the less reliable such estimators will become.

2 Nonparametric Deconvolution

An interesting special case of the unfolding problem is obtained when attention is restricted to the case $g(X, \varepsilon) = X + \varepsilon$. The integral equation relating the measured and true density reduces to a convolution equation

$$f_Y(y) = \int_{\mathbb{R}} f_\varepsilon(y - x) f_X(x) dx \Leftrightarrow f_Y = f_\varepsilon * f_X.$$

We will concentrate on this special problem in the nonparametric case as, on the one hand, it contains the germs of generality, while on the other hand, it is very well understood, having been extensively studied in the statistics literature. Our presentation in this section will borrow heavily from Meister [9], who provides an elegant overview of statistical deconvolution, and where precise versions of the statements given here, along with proofs may be found.

2.1 Inversion and Ill-Posedness

The fact that the integral equation relating the two densities is a convolution, immediately suggests an estimation technique based on direct inversion of the convolution operator:

1. Let ϕ_X, ϕ_ε, ϕ_Y be the characteristic functions (Fourier transforms) corresponding to the densities f_X, f_ε and f_Y, respectively.
2. Then $f_Y = f_\varepsilon * f_X \implies \phi_Y = \phi_X \phi_\varepsilon$.
3. So if $\hat{\phi}_Y$ is an estimate of the characteristic function of Y, we could estimate ϕ_X by $\hat{\phi}_Y / \phi_\varepsilon$, to obtain:

$$\hat{f}_X(t) = \frac{1}{2\pi} \int_{\mathbb{R}} e^{-itu} \frac{\hat{\phi}_Y(u)}{\phi_\varepsilon(u)} du.$$

This, of course, raises the question of how can one estimate the characteristic function ϕ_Y. Provided that approximation error is measured in square integrated distance, the answer is provided by the Plancherel identity: good estimates of f_Y give good estimates of ϕ_Y (and vice versa) by Fourier transforming,

$$\int [f_1(y) - f_2(y)]^2 dy = \int \left| \widetilde{f_1}(u) - \widetilde{f_2}(u) \right|^2 du.$$

Here $\widetilde{g}$ denotes the Fourier transform of g. So if $\hat{f}_Y$ is a good estimator of f_Y, then $\widetilde{\hat{f}}_Y$ is just as good an estimator of ϕ_Y, so that we may estimate f_Y, and then apply the Fourier transform to get an estimator of ϕ_Y. However, the continuity (boundedness) of the operation $f_X \mapsto f_X * f_\varepsilon$ (and corresponding discontinuity (unboundedness) of the inverse operation) will reveal that it is not enough to try to estimate f_Y (or, equivalently ϕ_Y) accurately.

Observe that, by the Plancherel identity, we have

$$\|f_X - \hat{f}_X\|^2 = \int |f_X(x) - \hat{f}_X(x)|^2 dx = \int \left| \phi_X(u) - \frac{\hat{\phi}_Y(u)}{\phi_\varepsilon(u)} \right|^2 du = \int \left| \frac{\phi_Y(u) - \hat{\phi}_Y(u)}{\phi_\varepsilon(u)} \right|^2 du.$$

where the last equality follows from $f_Y = f_X * f_\varepsilon$. Now, typically, $\phi_\varepsilon(u)$ decays "fast" as $u \to \pm\infty$. Therefore, a small discrepancy between $\hat{\phi}_Y$ and ϕ_Y for a large frequency u, can be blown up to an arbitrarily large discrepancy between f_X and $\hat{f}_X$. Such discrepancies are guaranteed to occur by the inherent statistical variability, but also because of sample limitations: it is intuitively clear that estimating the highest frequency characteristics of f_Y accurately based on a finite sample is essentially not possible.

2.2 Regularisation

We have seen that at the essence of ill-posedness lies the fact that $\frac{\hat{\phi}_Y(u)}{\phi_\varepsilon(u)}$ is potentially a bad estimator of $\phi_X(u)$ for u outside some bounded interval. Nevertheless, if f_X is going to be square integrable, we expect $|\varphi_X(u)|$ to become negligible as $u \to \pm\infty$. That is, we might a priori know based on qualitative properties of f_X that ϕ_X is practically zero outside some domain $\left[-\frac{1}{b}, \frac{1}{b}\right]$. Given this information, we can set our estimator to be $\frac{\hat{\phi}_Y(u)}{\phi_\varepsilon(u)}$ when $u \in \left[-\frac{1}{b}, \frac{1}{b}\right]$, and set it to zero outside that domain to avoid the blow-up. That is, we employ the *regularised* estimator

$$\frac{\mathbf{1}\{-1/b \le u \le 1/b\}\hat{\phi}_Y(u)}{\phi_\varepsilon(u)} = \frac{\mathbf{1}\{-1 \le bu \le 1\}\hat{\phi}_Y(u)}{\phi_\varepsilon(u)} = \frac{\widetilde{K}(bu)\hat{\phi}_Y(u)}{\phi_\varepsilon(u)},$$

where we used the notation $\widetilde{K}(bu) = \mathbf{1}\{-1 \le bu \le 1\}$ to imply that we understand $\mathbf{1}\{-1 \le bu \le 1\}$ as the Fourier transform of some *kernel function* K. More generally, we could use a weight function $\widetilde{K}(bu) := \widetilde{K}_b(u)$, where $\widetilde{K}(u)$ is supported on $[-1, 1]$ and bounded. This corresponds to estimating ϕ_Y by $\widetilde{K}(bu)\hat{\phi}_Y(u)$ which leaves the naive estimate unaffected for small frequencies, tames the variation for higher frequencies, and kills the variation completely beyond a certain frequency threshold.

2.3 The Naive and Kernel Estimators

To illustrate these ideas, we consider the so-called 'plug-in' estimator of ϕ_Y, defined as

$$\hat{\phi}_Y^{nve}(u) = \frac{1}{n}\sum_{j=1}^{n} e^{iuY_j}.$$

We labeled this as the naive estimator ('*nve*'), since it merely uses the empirical version of the characteristic function. It is, nevertheless, a reasonable estimator for every fixed u by the strong law of large numbers: it is unbiased, $\mathbb{E}[\hat{\phi}_Y^{nve}(u)] = \phi_Y(u)$, and for any u, $\mathbb{E}|\hat{\phi}_Y^{nve}(u) - \phi_Y(u)|^2 = O(n^{-1})$. However, these are properties that hold locally, i.e. for a fixed u. When employing $\hat{\phi}_Y^{nve}$ as part of a deconvolution estimator to estimate f_X, we see that it is "totally affected" by ill-posedness: the induced estimator $\hat{f}_X^{nve}(x) = \text{“}(2\pi)^{-1}\sum_{j=1}^{n}\int \frac{e^{iu(Y_j - x)}}{n\phi_\varepsilon(u)}du\text{”}$ is not even well-defined (hence the quotation marks – the quantity inside the integral is not integrable). Nevertheless, we may employ the regularisation strategy presented in the previous section, in order to obtain a regularised version of the naive estimator that is not only well-defined, but also has controllable variation:

$$\hat{f}_X^{kernel}(x) = \frac{1}{2\pi}\int e^{-iux}\frac{\widetilde{K}(bu)\sum_{j=1}^{n}\frac{1}{n}e^{iuY_j}}{\phi_\varepsilon(u)}du.$$

We call this the kernel estimator, because we have built it by dampening the high frequency components of the naive estimator using the Fourier transform of some kernel K.

2.4 Error Properties of the Kernel Estimator

Since the object being estimated is a function, it is not immediately clear what criterion one might employ in order to measure the accuracy of the deconvolution estimator. Natural building blocks for error measures are divergences on function space, which can then yield error measures by means of averaging (averaging meaning taking the expectation with respect to the sample observations $Y_1, ..., Y_n$). The choice of divergence reflects which aspects of f_X we wish to emphasize the most. For example, one could define an error measure based on the Cramér-Von Mises divergence as $\mathbb{E}[\int(\hat{F}_X(x) - F_X(x))^2 f_X(x)dx]$, placing greater emphasis on regions of high density. If interest lies primarily on the tails of the distribution, then one could employ an Anderson-Darling divergence, $\mathbb{E}[\int(\hat{F}_X(x) - F_X(x))^2[F_X(x)(1 -$

$F_X(x))]^{-1} f_X(x)dx]$. Perhaps the most widely studied error measure is the mean integrated squared error measure, $\mathbb{E}\int(\hat{f}_X(x) - f_X(x))^2 dx = \mathbb{E}\|\hat{f}_X - f_X\|^2$, which places equal emphasis on different parts of the domain of f_X. In what follows, we will concentrate on this particular error measure. This is also partly a matter of convenience, since the convolution operator is naturally linked with Fourier analysis on L^2. Assuming that f_X is square-integrable and that $\phi_\varepsilon(u) \neq 0$ everywhere, then, if $\widetilde{K}(u)$ is bounded and supported on a bounded interval, it can be shown that

$$\mathbb{E}\|f_x - \hat{f}_X^{kernel}\|^2 = \frac{1}{2\pi n}\int |\widetilde{K}(bu)|^2 \left[\frac{1}{|\phi_\varepsilon(u)|^2} - |\phi_X(u)|^2\right] du + \frac{1}{2\pi}\int |\widetilde{K}(bu) - 1|^2 |\phi_X(u)|^2 du.$$

This error expression provides insight into the nature of the estimation error in deconvolution. In the statistical terminology, the first term represents the variance component of the error, whereas the second term represents the bias component. The variance term represents the component of the error that is due to statistical variation as well as the instability of the inversion process. The bias component describes the systematic error due to regularisation. The expression reveals the existence of a fundamental trade-off between the two terms, governed by the *regularisation parameter* b:

1. In the first term (variance term), $\widetilde{K}$ controls blow-up caused by rapid decay of ϕ_ε. This component is decreasing in b (i.e. decreases as the length of the interval $[-b^{-1}, b^{-1}]$ decreases).
2. In the second term (bias term), $\widetilde{K}$ controls "how far we are on average" from estimating ϕ_X. This component is increasing in b (i.e. increases as the length of the interval $[-b^{-1}, b^{-1}]$ decreases).

Therefore, the choice of b must be made judiciously, in order to balance these two effects, and minimise the overall mean squared error.

One may, however, pose the question of whether it is possible to do better than the kernel estimator in terms of overall error. Said differently, is the kernel estimator "optimal" (or at least fairly reasonable), or should we rather concentrate on something different? As is typically the case in statistics, this question cannot be answered exactly, i.e. for fixed sample size, nor in complete generality (very weak specification of the properties of the densities involved). A partial answer can be given in an asymptotic regime, as the sample size is taken to increase to infinity, $n \to \infty$, and under a stronger specification of the function class of the densities involved and the properties of the error density. Roughly speaking, we could require that the class $\mathcal{F}$ contains relatively smooth functions (e.g. densities possessing a β-th derivative with uniformly bounded L^2 norm) and that the tail decay of the error characteristic function is of polynomial order, say with an exponent $\alpha < 0$ (such error densities are called smooth, to be contrasted with supersmooth error densities, where the decay of the characteristic function is exponential). Notice that the latter assumption is related to the typical magnitude of the errors: the rate of decay of the error characteristic function is connected with the typical magnitude of the error (slow tail decay of the characteristic function means that the error density is concentrated around zero).

Under these assumptions, and if the kernel K satisfies certain regularity conditions, choosing $b \sim n^{-\frac{1}{2\beta+2\alpha+1}}$ as $n \to \infty$, we can obtain the following asymptotic upper bound for the error of the kernel estimator:

$$\sup_{f\in\mathcal{F}} \|\hat{f}_X^{kernel} - f\|^2 = O(n^{-\frac{2\beta}{2\beta+2\alpha+1}}).$$

On the other hand, if we additionally assume that $|\frac{d}{du}\phi_\varepsilon(u)| \leq \text{const} \times |t|^{-\alpha}$, and $\beta > 1/2$, we have

$$\sup_{f\in\mathcal{F}} \|\hat{f}_X - f\|^2 \geq \text{const} \times n^{-\frac{2\beta}{2\beta+2\alpha+1}},$$

for any estimator $\hat{f}_X$ of f_X, as $n \to \infty$ (the precise regularity conditions can be found in Meister [9, Sec. 2.4]). That is, asymptotically, we will never be able to perform order of magnitude better than the kernel estimator over the whole function class, since the upper bound for the kernel estimator coincides up to

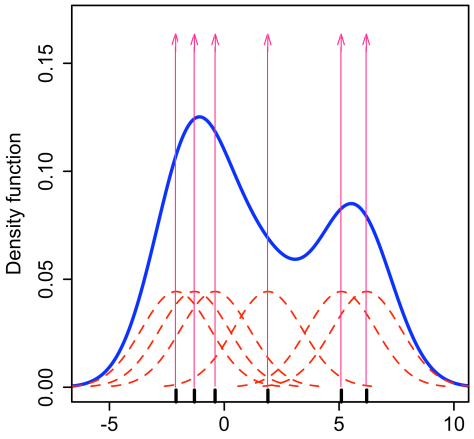

Fig. 2: Schematic representation of the workings of a kernel estimator in the time domain.

a constant with the lower bound for *any* other estimator. Of course these results leave much room for discussion, as they are uninformative when it comes to the exact finite sample behaviour of the kernel estimator relative to other estimators. In addition, these results show rate optimality, but there is an undetermined constant involved. Finally note that they consider the worst case error over the function class. They should therefore be interpreted with care as qualitative statements. Optimality results of a similar flavour are available when the error density is super-smooth (e.g. Gaussian), but with slower convergence rates.

2.5 Time Domain Interpretation of the Kernel Estimator

The kernel estimator of ϕ_Y amounts to replacing the naive estimator, $\frac{1}{n}\sum_{j=1}^{n} e^{iuY_j}$, by the regularised estimator $\widetilde{K}(bu) \cdot \frac{1}{n}\sum_{j=1}^{n} e^{iuY_j}$. Applying the inverse Fourier transform, we see that this amounts to estimating f_Y by

$$\left[\frac{1}{b}K\left(\frac{x}{b}\right)\right] * \left[\frac{1}{n}\sum_{j=1}^{n}\delta(Y_j - x)\right] = \sum_{j=1}^{n}\frac{1}{nb}K\left(\frac{Y_j - x}{b}\right),$$

where $\widetilde{K}(u) = \int K(t)e^{itu}dt$ and δ is Dirac's delta. Intuitively, instead of using the "empirical density" as the estimator of f_Y, the kernel estimator uses a smoothed version. In a sense, the empirical density 'fits best' to the observed data but is far too rough (it is not even a function in a proper sense). This roughness of the density estimate would translate into a slow tail decay of its Fourier transform, leading to the problems observed in Section (2.3). This is the essence of the bias-variance trade-off in the time domain: we need to attempt to fit the data well but at the same time maintain a certain level of smoothness to avoid ill-posedness issues.

2.6 Tuning The Regularisation Parameter

The bias-variance tradeoff phenomenon requires that we tune the regularisation parameter in order to obtain the optimal amount of regularisation. For supersmooth error densities (e.g. Gaussian), the (asymptotically) optimal regularisation parameter is independent of the unknown function f_X, and depends only on the error density – in fact, provided that the error density is known, we can determine the exact value of the regularisation parameter (there is no unknown constant involved).

However, in the case of simply smooth error density [assumption (A3)] we saw that the bandwidth depends on the smoothness properties of the unknown f_X. In fact, even if we knew these exactly, we

would still only know how to choose a bandwidth up to a constant. An alternative is to choose b using the data as a guide. For example, we could attempt to choose b to minimise $\|\hat{f}_{b,X}^{kernel} - f_X\|^2$. This, of course, depends on the unknown density f_X, but can be estimated using the data. The squared norm admits the decomposition

$$\|\hat{f}_{b,X}^{kernel} - f_X\|^2 = \|\hat{f}_{b,X}^{kernel}\|^2 - 2\Re\underbrace{\left[\langle \hat{f}_{b,X}^{kernel}, f_X\rangle\right]}_{\mathbb{E}\left[\hat{f}_{b,X}^{kernel}(X)\right]} + \|f_X\|^2,$$

where $\Re$ denotes the real part of a complex number. We notice that the first term depends only on the data, the second term is estimable from the data and the third term is independent of b. One could employ leave-one-out cross validation in the Fourier domain to estimate $\mathbb{E}\left[\hat{f}_{b,X}^{kernel}(X)\right]$ and select the value of b that minimises the overall expression (see Meister [9, Sec. 2.5.1] for details).

The problem with cross-validation is that b becomes a random variable, dependent on $\hat{f}_{b,X}^{kernel}$. In essence, we use the data twice and so our error and optimality results are not guaranteed to hold true. Nevertheless, at least asymptotically, cross validation can be seen to be adaptive – meaning that with increasing sample size, it will eventually provide the optimal error rates, thus *adapting* to the potentially unknown smoothness class of the unknown density. Under some additional smoothness assumptions on f_X (which we omit for brevity), as well as the α-polynomial tail decay assumption on ϕ_ε, if we use the sinc kernel to construct $\hat{f}_{b,X}^{kernel}$ and optimise the empirical integrated squared error on a fine enough grid $G(n,\alpha)$ (depending on n and α), we obtain:

$$\limsup_{n\to\infty}\left\{\frac{\inf\limits_{b\in G(n,\alpha)}\widehat{\mathrm{ISE}}(b)}{\inf\limits_{b>0}\mathbb{E}\|\hat{f}_{b,X}^{kernel} - f_X\|^2}\right\} \leq 1, \qquad \text{with probability 1.}$$

Here, $\widehat{\mathrm{ISE}}(b)$ denotes the cross-validated mean integrated squared error corresponding to a regularisation parameter b. See Meister [9, Thm. 2.17] for the precise statement.

2.7 A Pointwise Central Limit Theorem

One may be interested to obtain a confidence interval for the value of the unknown density estimate at a point x. For this reason, one would require an approximate distribution for $\hat{f}_X$ at the point x. If $f_Y = f_X * f_\varepsilon$ is uniformly bounded, then, under regularity conditions on the kernel and polynomial decay of the error characteristic function (smooth error density),

$$\frac{\hat{f}_X^{kernel}(x) - \mathbb{E}[\hat{f}_X^{kernel}(x)]}{\sqrt{\mathrm{Var}[\hat{f}_X^{kernel}(x)]}} \xrightarrow{d} N(0,1),$$

for all $x \in \mathbb{R}$, provided that b is selected so that $bn \overset{n\to\infty}{\longrightarrow} \infty$. This central limit theorem can be used in conjunction with the available bounds on the squared bias in order to construct asymptotic confidence intervals.

It should be noted that there is no corresponding central limit theorem when the error density is supersmooth (exponential decay of the characteristic function), e.g. in the case of Gaussian errors.

2.8 Induced Estimators of the Distribution Function

It might be the case that we are not interested in estimating the density f_X per se, but rather that we are interested in estimating the probability of a certain interval $[c,d]$,

$$F_{c,d} = F_X(d) - F_X(c) = \mathbb{P}[c \leq X \leq d] = \int_c^d f_X(x)dx.$$

Here F_X denotes the distribution function of the random variable X. From the mathematical point of view, this is an 'easier' problem, as we are required to estimate a smooth functional of the density function. Since we have already done the hard work of estimating f_X, it is natural to use a plug-in estimator for this purpose, i.e. use the estimator

$$\hat{F}_{c,d} = \int_a^b \hat{f}_X^{kernel}(x)dx = \int \mathbf{1}\{c \le x \le d\}\hat{f}_X^{kernel}(x)dx.$$

Plancherel's identity now shows that

$$\hat{F}_{c,d} = \frac{1}{n\pi}\sum_{j=1}^{n}\int e^{iu(c+d)/2}\sin[(u(d-c)/2]\frac{\widetilde{K}(bu)e^{iuY_j}}{u\phi_\varepsilon(u)}du.$$

Under some smoothness assumptions on the unknown density function, we can also obtain error bounds for this estimator of the distribution function, both for a fixed unknown density, as well as uniformly in a prescribed function class. In particular, we suppose that the class $\mathcal{F}$ contains uniformly bounded and smooth densities (possessing a globally Hölder-continuous $\lfloor\beta\rfloor$-th derivative, $\lfloor\beta\rfloor$ denoting integer part the integer part of β). Then, provided the kernel satisfies certain regularity conditions, the plug-in estimator satisfies

$$\mathbb{E}|\hat{F}_{c,d} - F_{c,d}|^2 \le \|f_Y\|_\infty \frac{2}{\pi n}\int \frac{\left|\sin[u(d-c)/2]\widetilde{K}(bu)\right|^2}{|u|^2|\phi_\varepsilon(u)|^2}du + 4C^2b^2\left(\int |K(t)||t|dt\right)^2.$$

Furthermore, if we additionally assume an α-polynomial decay of ϕ_ε, we may also obtain the following uniform error bounds for $\sup_{f\in\mathcal{F}}\mathbb{E}|\hat{F}_{c,d} - F_{c,d}|^2$:

1. For $0 < \alpha < 1/2$, if we simply put $\widetilde{K} = 1$, the bound is $O(1/n)$.
2. For $\alpha = 1/2$ and $b \sim n^{-1}$, the bound is $O\left(\frac{\log n}{n}\right)$.
3. For $\alpha > 1/2$ and $b \sim n^{-1/(2\beta+2\alpha+1)}$, the bound is $O\left(n^{-(2\beta+2)/(2\beta+2\alpha+1)}\right)$.

Notice that for $0 < \alpha < 1/2$ (slow tail decay of the error characteristic function), the rate is essentially the parametric estimation rate. The precise regularity conditions may be found in Meister [9, Sec. 2.7].

3 Nonparametric Unfolding

We now turn to making some broad remarks on a more general version of unfolding. This arises when considering a more general binary operation $g(X, \varepsilon)$. The integral equation now becomes

$$f_Y(y) = \int h(x,y)f_X(x)dx \implies f_Y = \mathcal{L}f_X,$$

where $\mathcal{L} : f \mapsto \int h(x,y)f(y)dy$ is a more general integral operator. For simplicity, we assume that the densities involved are square integrable and supported on $[a,b]$, and that $h(x,y) = \sum_{k=1}^{\infty}\lambda_k\varphi_k(x)\varphi_k(y)$ for an orthonormal basis $\{\varphi_k\}$ of $L^2[a,b]$ and a square summable sequence of non-zero real coefficients $\{\lambda_k\}$ (so that h is square-integrable and symmetric). Then, the eigenfunctions of the operator $\mathcal{L}$ are precisely the φ_k, with corresponding eigenvalues λ_k, $\mathcal{L}\varphi_k = \lambda_k\varphi_k$. We have $f(x) = \sum_{k=1}^{\infty}a_k\phi_k$ for all square integrable f, with $a_k = \langle f, \varphi_k\rangle = \int_a^b f(t)\varphi_k(t)dt$. In this setting, we may write

$$f_Y = \mathcal{L}f_X = \mathcal{L}\left[\sum_{k=1}^{\infty}\langle f_X, \varphi_k\rangle\varphi_k\right] = \sum_{k=1}^{\infty}\langle f_X, \varphi_k\rangle\mathcal{L}[\varphi_k] = \sum_{k=1}^{\infty}\lambda_k\langle f_X, \varphi_k\rangle\varphi_k.$$

On the other hand, we must also have $f_Y = \sum_{k=1}^{\infty} \langle f_Y, \varphi_k \rangle \varphi_k$ so that $\langle f_Y, \varphi_k \rangle = \lambda \langle f_X, \varphi_k \rangle \varphi_k$, since the Fourier representation is unique. Going backwards, we can invert the transform by taking $f_X = \sum_{k=1}^{\infty} \frac{\langle f_Y, \varphi_k \rangle}{\lambda_k} \varphi_k$. This short analysis suggests an estimation strategy similar to that employed in the case of deconvolution: (1) First estimate f_Y by some reasonable estimator $\hat{f}_Y$, (2) Then apply the inverse transform to $\hat{f}_Y$,

$$\hat{f}_X = \sum_{k=1}^{\infty} \frac{\langle \hat{f}_Y, \varphi_k \rangle}{\lambda_k} \varphi_k.$$

We now observe how ill-posedness manifests itself in this context: since the sequence of eigenvalues is assumed square integrable, it must be that $\lambda_k \downarrow 0$. Note that since $\{\varphi_k\}$ is a basis,

$$\|f_X - \hat{f}_X\|^2 = \sum_{k=1}^{\infty} \left(\langle f_X, \varphi_k \rangle - \frac{\langle \hat{f}_Y, \varphi_k \rangle}{\lambda_k} \right)^2 = \sum_{k=1}^{\infty} \frac{\left(\langle f_Y, \varphi_k \rangle - \langle \hat{f}_Y, \varphi_k \rangle \right)^2}{\lambda_k^2}.$$

The situation is thus similar to the deconvolution setting: small estimation errors in the estimation of f_Y (e.g. in the estimation of $\langle f_Y, \varphi_k \rangle$ for large k) magnify to large estimation errors of f_X due to the blow-up of the inverse eigenvalues. As in deconvolution, we will need to 'tame' this blow-up by means of regularisation. To this aim, we can argue that since $\|f_X\|^2 = \sum_{k=1}^{\infty} \langle f_X, \varphi_k \rangle^2 < \infty$, it must be that $\langle f_X, \varphi_k \rangle \to 0$. Therefore, we may choose a truncation level B (regularisation parameter), and enforce:

$$\langle \hat{f}_X, \varphi_k \rangle = \begin{cases} \frac{\langle \hat{f}_Y, \varphi_k \rangle}{\lambda_k} & \text{if } k \leq B, \\ 0 & \text{otherwise.} \end{cases}$$

Equivalently, we estimate f_Y by $\sum_{k=1}^{B} \langle \hat{f}_Y, \varphi_k \rangle \varphi_k$, obtaining the estimator of the original density:

$$\hat{f}_X = \sum_{k=1}^{B} \frac{\langle \hat{f}_Y, \varphi_k \rangle}{\lambda_k} \varphi_k.$$

To illustrate the approach, we apply this *spectral truncation regularisation* to a naive 'plug-in' estimator of f_Y. This is obtained by noting that $\langle f_Y, \varphi_k \rangle = \int_a^b \varphi_k(u) f_Y(u) du = \mathbb{E}[\varphi_k(Y)]$. Consequently $f_Y(u) = \sum_{k=1}^{\infty} \mathbb{E}[\varphi_k(Y)] \varphi_k(u)$. Again, the law of large numbers seems to suggest to estimate the expectations $\mathbb{E}[\varphi_k(Y)]$ by their empirical versions (their optimal unbiased estimators) to obtain the naive estimator

$$\hat{f}_Y^{nve}(u) = \text{“}\sum_{k=1}^{\infty} \left[\frac{1}{n} \sum_{j=1}^{n} \varphi_k(Y_j) \right] \varphi_k(u)\text{”}.$$

Just as with deconvolution, though, this is not even a well-defined estimator. To see why, notice that the series would appear to "converge" to the "Fourier series expansion" of the empirical density $\frac{1}{n} \sum_{j=1}^{n} \delta(u - Y_j)$ with respect to the $\{\varphi_k\}$ basis, since $\frac{1}{n} \sum_{j=1}^{n} \varphi_k(Y_j) = \int \left[\frac{1}{n} \sum_{j=1}^{n} \delta(u - Y_j) \right] \varphi_k(u) du$. But this empirical density is far from being square integrable, and such an expansion is undefined. Nevertheless, any finite sum formed by the first B summands of the series is well defined, that is, the spectrally truncated naive estimator $\hat{f}_Y^{trunc}$, is well-defined, yielding the truncated series estimator of f_X,

$$\hat{f}_X^{trunc}(x) = \sum_{k=1}^{B} \left[\frac{1}{n\lambda_k} \sum_{j=1}^{n} \varphi_k(Y_j) \right] \varphi_k(x).$$

In terms of error properties, one can see that the truncated estimator satisfies a similar bias-variance tradeoff as before,

$$\mathbb{E}\|f_X - \hat{f}_X^{trunc}\|^2 = \sum_{k=1}^{B} \frac{\text{Var}[\varphi_k(Y)]}{n\lambda_k^2} + \sum_{k=B}^{\infty} \langle f_X, \varphi_k \rangle^2.$$

In the first term (variance term), B controls the blow-up caused by the decay of eigenvalues as compared to the behaviour of $\text{Var}[\varphi_k(Y)]/n$. In the second term (bias term), B controls our systematic deviation from f_X, expressed in terms of what part of f_X we are missing completely due to the truncation.

The spectrally truncated naive estimator also allows one to appreciate the potential effects that dependence among the measurement error inputs ε_i may bring about: if the Y_i are stationary but not independent, then we can still rely on $n^{-1}\sum_{i=1}^n \varphi_k(Y_i)$ as an estimator of $\mathbb{E}[\varphi_k(Y)]$ by the ergodic theorem, but the quality of the estimator for finite n may suffer, depending on the strength of the dependence (the bias term will remain the same, but the terms in the series of the variance component will now be $(n\lambda_k)^{-2}\sum_{i=1}^n\sum_{j=1}^n \text{Cov}(\varphi_k(Y_i),\varphi_k(Y_j))$).

We now turn to show that, in fact, the spectrally truncated estimator admits a kernel estimator interpretation. Let $K_B(x,y)=\sum_{k=1}^B \varphi_k(x)\varphi_k(y)$, and observe that

$$\hat{f}_Y^{trunc}(x)=\sum_{k=1}^B\langle \hat{f}_Y^{nve},\varphi_k\rangle\varphi_k(x)=\sum_{k=1}^B\varphi_k(x)\int_a^b\varphi_k(y)\hat{f}_Y^{nve}(y)dy=\int_a^b K_B(x,y)\hat{f}_Y^{nve}(y)dy.$$

We conclude this brief discussion of the more general unfolding framework, by pointing out a slightly different observation scenario that one may consider in practice. Instead of observing an iid sample from f_Y, we might assume that we observe f_Y itself, subject to some error, $f_Y=\mathcal{L}f_X+\epsilon$. For example, ϵ can be thought of as white noise, meaning that we are able to observe

$$\langle f_Y,\varphi_k\rangle=\lambda_k\langle f_X,\varphi_k\rangle+\langle\epsilon,\varphi_k\rangle,\qquad k=1,2,...$$

with $\langle\epsilon,\varphi_k\rangle$ an iid white Gaussian noise sequence. This point of view would lead to the more classical statistical inverse problem framework, which is very well understood (see Cavalier [3]). Similar considerations apply, with spectral truncation and Tikhonov regularisation being the main approaches to relax the ill-posed problem. Variants of cross validation, or other methods for the tuning of the amount of regularisation such as Stein risk and the risk hull method have been studied in this context (see, Cavalier & Golubev [4]).

4 Discussion and Further Details

A question that took up a significant part of the discussion at the end of the unfolding session was, plainly stated, “to fold or to unfold"? In particular, should one attempt to estimate the folded function in Y-space and then unfold (invert the integral transform) in order to obtain their estimate in X-space, or rather should one look for the density in X-space such that, when folded (pushed forward through the integral transform), it would yield a density in Y-space that is most consistent with the data. From the mathematical point of view, the two views are essentially equivalent. If one chooses to fold instead of unfold, then one still needs to apply the same sort of regularisation: functions in X-space that are significantly far apart, may yield almost the same folded density in Y-space. To see this, consider two densities in X-space whose Fourier coefficients with respect to the first B eigenfunctions of the folding operator are identical, but the remaining coefficients are significantly different, though not different enough to counterbalance the decay of the eigenvalues of the operator. In this case, regularisation would amount to restricting one’s search on a subspace of X-space spanned by the first B eigenfunctions of the folding operator (where B would be a regularisation parameter to be tuned judiciously). Either approach could therefore be adopted, depending on what is most convenient from a practical point of view – but folding *does not circumvent* the problem of ill-posedness if regularisation is not applied.

Among the points raised was the use of cross-validation to select a regularisation parameter. The issue was connected to the feasibility of conducting the leave-one-out cross validation given that the sample size may be of the order of hundreds of thousands of observations. This, however, is not necessarily a problem: leave-one-out cross validation is employed in situations where the sample size is relatively

small, in order to avoid splitting the sample into two parts (a validation and an estimation part). For very large samples, one could employ an approach that is less computationally intensive, by splitting the data.

Another issue was that the folding operator is often only approximately known. This will indeed perturb things, but the main principles remain the same. For example, a simple approach to inject randomness into the operator in the setting of Section 3 would be by assuming that the eigenvalues are in fact a random element in the space of square summable sequences. This would not change the approach fundamentally. More complicated scenarios could introduce random eigenfunctions.

Below we provide a very short reference list on some of the topics covered in this overview. Silverman [10] provides an accessible introduction to nonparametric density estimation, while Meister [9] contains an elegant overview of the statistical deconvolution problem. One of the earliest papers in statistical deconvolution is Stefanski & Carroll [11], while a treatment of convergence rates can be found in Carroll & Hall [2] and Fan [7]. Details on bandwidth selection can be found in Delaigle & Hall [6], while Bissantz et al. [1] study the problem of constucting confidence bands for deconvolved density estimates. Hall & Lahiri [8] consider the problem of distribution estimation in the deconvolution setting. Finally, Cavalier [3] provides a review of some of the basic aspects of more general statistical inverse problems –such as the unfolding problem– in the context provided in the end of Section 3; in the same context, Cavalier & Hengartner [5] consider the case where the eigenvalues of the folding operator are noisy.

Acknowledgement

I wish to thank David Cox for kindly reading through a draft version of this report and providing useful comments. My thanks also go to Louis Lyons, for providing suggestions that helped improve the presentation.

References

[1] Bissantz, N., Dümbgen, L., Holzmann, H. & Munk, A. (2007). Nonparametric Confidence bands in deconvolution density estimation. *J. Roy. Stat. Soc. Ser. B,* ***69****: 483–506.*

[2] Carroll, R.J. & Hall, P. (1988). Optimal rates of convergence for deconvolving a density. *J. Am. Statist. Assoc.,* ***83****: 1184–1186.*

[3] Cavalier, L. (2008). Nonparametric statistical inverse problems. *Inverse Problems,* ***24****: 034004.*

[4] Cavalier, L. & Golubev, Yu. (2006). Risk hull method and regularization by projections of ill-posed inverse problems. *Annals of Statistics,* ***34****: 1653–1677.*

[5] Cavalier, L. & Hengartner, N.W. (2005). Adaptive estimation for inverse problems with noisy operators. *Inverse Problems,* ***21****: 1345.*

[6] Delaigle, A. & Hall, P. (2006). On the optimal kernel choice for deconvolution. *Stat. Prob. Lett.,* ***76****: 1594–1602.*

[7] Fan, J. (1991). On the optimal rates of convergence for non-parametric deconvolution problems. *Ann. Stat.* ***19****: 1257–1272.*

[8] Hall, P. & Lahiri, S.N.(2008). Estimation of distributions, moments and quantiles in deconvolution problems. *Ann. Stat.* ***36****: 2110–2134.*

[9] Meister, A. (2009). *Deconvolution Problems in Nonparametric Statistics*. Springer.

[10] Silverman, B.W. (1998). *Density Estimation*. Chapman & Hall.

[11] Stefanski, L.A. & Carroll, R.J. (1990). Deconvoluting kernel density estimators. *Statistics,* ***20****: 169–184.*