To Appear: Neural Computation, February 2011 (in press).

Least squares estimation without priors or supervision

Martin Raphan and Eero P. Simoncelli

Howard Hughes Medical Institute,
Center for Neural Science, and
Courant Inst. of Mathematical Sciences
New York University

Email: *raphan@cims.nyu.edu, eero.simoncelli@nyu.edu*

8 November 2010

Abstract

Selection of an optimal estimator typically relies on either supervised training samples (pairs of measurements and their associated true values), or a prior probability model for the true values. Here, we consider the problem of obtaining a least-squares estimator given a measurement process with known statistics (i.e., a likelihood function), and a set of *unsupervised* measurements, each arising from a corresponding true value drawn randomly from an *unknown* distribution. We develop a general expression for a nonparametric empirical Bayes least squares (NEBLS) estimator, that expresses the optimal least-squares estimator in terms of the measurement density, with no explicit reference to the unknown (prior) density. We study the conditions under which such estimators exist, and derive specific forms for a variety of different measurement processes. We further show that each of these NEBLS estimators may be used to express the mean squared estimation error as an expectation over the measurement density alone, thus generalizing Stein's unbiased risk estimator (SURE) which provides such an expression for the additive Gaussian noise case. This error expression may then be optimized over noisy measurement samples, in the absence of supervised training data, yielding a generalized SURE-optimized parametric least squares (SURE2PLS) estimator. In the special case of a linear parameterization (i.e., a sum of nonlinear kernel functions), the objective function is quadratic, and we derive an incremental form for learning this estimator from data. We also show that combining the NEBLS form with its corresponding generalized SURE expression produces a generalization of the "score matching" procedure for parametric density estimation. Finally, we have implemented several examples of such estimators, and we show that their performance is comparable to their optimal Bayesian or supervised regression counterparts for moderate to large amounts of data.

Keywords: least squares estimation, Bayesian estimation, nonparametric empirical Bayes least squares, generalized SURE parametric least squares, SUREshrink, score matching, prior-free estimation, unsupervised regression

1 Introduction

The problem of estimating signals based on partial, corrupted measurements arises whenever a machine or biological organism interacts with an environment that it observes through sensors. Optimal estimation has a long history, documented in the published literature of a variety of communities: statistics, signal processing, sensory perception, motor control, and machine learning, just to name a few. The two most commonly used methods of obtaining an optimal estimator are (1) Bayesian inference, in which the estimator is chosen to minimize the expected error over a posterior distribution, obtained by combining a prior distribution with a model of the measurement process; and (2) supervised regression, in which the estimator is selected from a parametric family by minimizing the error over a training set containing pairs of corrupted measurements and their correct values. Here, we consider the problem of obtaining a least-squares (also known as minimum mean squared error) estimator, in the absence of either supervised training examples or a prior model. Specifically, we assume a known measurement process (i.e., the probability distribution of measurements conditioned on true values), and we assume that we have access to a large set of measurement samples, each arising from a corresponding true value drawn from an unknown signal distribution. The goal, then, is to obtain a least-squares estimator given these two ingredients.

The statistics literature describes a framework for handling this situation, generally known as *Empirical Bayes* estimation, in which the estimator is learned from observed (noisy) samples. More specifically, in the *parametric* Empirical Bayes approach, one typically assumes a parametric form for the density of the clean signal, and then learns the parameters of this density from the noisy observations. This is usually achieved by maximizing likelihood or matching moments (Morris, 1983; Casella, 1985; Berger, 1985), both of which are inconsistent with our goal of minimizing mean squared estimation error. In addition, if one assumes a parametric form for the density of the clean signal which can not give a good fit to the true distribution, performance is likely to suffer. For the case of Poisson observations, Robbins introduced a *nonparametric* Empirical Bayesian *Least Squares* (NEBLS) estimator that is expressed directly in terms of the measurement density, without explicit reference to the prior (Robbins, 1956). This remarkable result was subsequently extended to cases of additive Gaussian noise (Miyasawa, 1961), and to general exponential measurements (Maritz and Lwin, 1989), where it was referred to as a *simple empirical Bayes* estimator.

Here, we develop a general expression for this type of NEBLS estimator, written in terms of a linear functional of the density of noisy measurements. The NEBLS estimators previously developed for Poisson, Gaussian, and exponential measurement processes are each special cases of this general form. We provide a complete characterization of observation models for which such an estimator exists, develop a methodology for obtaining the estimators in these cases, and derive specific solutions for a variety of corruption processes, including the general additive case and various scale mixtures. We also show that any of these NEBLS estimators can be used to derive an expression for the mean squared error (MSE) of an arbitrary estimator that is written as an expectation over the measurement density (again, without reference to the prior). These expressions provide a generalization of Stein's Unbiased Risk Estimate (SURE), which corresponds to the special case of additive Gaussian noise (Stein, 1981)), and analogous expressions that have been developed for the cases of continuous and discrete exponential families (Berger, 1980; Hwang, 1982). In addition to unifying and generalizing these previous examples, our derivation ties them directly to the seemingly unrelated NEBLS methodology.

In practice, approximating the reformulated MSE with a sample average allows one to optimize a *paramet-*

ric estimator based entirely on a set of corrupted measurements, without the need for the corresponding true values that would be used for regression. This has been done with SURE (Donoho and Johnstone, 1995; Pesquet and Leporini, 1997; Benazza-Benyahia and Pesquet, 2005; Luisier et al., 2006; Raphan and Simoncelli, 2007b; Blu and Luisier, 2007; Raphan and Simoncelli, 2008; Chaux et al., 2008), and recently with the analogous exponential case (Eldar, 2009)). We refer to this as a *generalized SURE-optimized parametric least squares* estimator (since the generalized SURE estimate is used to obtain the parameters of the least squares estimator, we use the acronym gSURE2PLS, with mnemonic pronunciation "generalized sure to please"). For the special case of an estimator parameterized as a linear combination of nonlinear kernel functions, we develop an incremental algorithm that simultaneously optimizes and applies the estimator to a stream of incoming data samples. We also show that the NEBLS solution may be combined with the generalized SURE expression to yield an objective function for fitting a parametric density to observed data, which provides a generalization of the recently developed "score matching" procedure (Hyvärinen, 2005, 2008). Finally, we compare the empirical convergence of several example gSURE2PLS estimators with that of their Bayesian counterparts. Preliminary versions of this work[1] have been presented in (Raphan and Simoncelli, 2007b; Raphan, 2007; Raphan and Simoncelli, 2009).

2 Introductory example: Additive Gaussian noise

We begin by illustrating the two forms of estimator for the scalar case of additive Gaussian noise (the vector case is derived in Sec. 6.1). Suppose random variable Y represents a noisy observation of an underlying random variable, X. It is well known that given a particular observation $Y = y$, the estimate that minimizes the expected squared error (sometimes called the *Bayes least squares* estimator) is the conditional mean:

$$\begin{aligned} \hat{x}(y) &= \mathbf{E}_{X|Y}\left(X|Y=y\right) \\ &= \int x P_{X|Y}(x|y)\,dx \\ &= \int x \frac{P_{X,Y}(x,y)}{P_Y(y)}\,dx, \end{aligned} \tag{1}$$

where the denominator contains the distribution of the observed data, which we refer to as the *measurement density* (sometimes called the *prior predictive density*). This can be obtained by marginalizing the joint density, which in turn can be written in terms of the prior on X using Bayes' Rule:

$$P_Y(y) = \int P_{X,Y}(x,y)\,dx = \int P_{Y|X}(y|x)\,P_X(x)\,dx. \tag{2}$$

2.1 Nonparametric empirical Bayes least squares estimator

The NEBLS estimation paradigm, in which the least squares estimator is written entirely in terms of the measurement density, was introduced by Robbins (1956), and was extended to the case of Gaussian additive noise by Miyasawa (1961). The derivation for the Gaussian case is relatively simple. First note that the

[1] In these previous publications, we referred to the generalized NEBLS estimator using the oxymoron "prior-free Bayesian estimator", and to the corresponding generalized SURE method as "unsupervised regression".

conditional density of the measurement given the true value is

$$P_{Y|X}(y|x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y-x)^2/2\sigma^2},$$

and thus, substituting into Eq. (2),

$$P_Y(y) = \int \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y-x)^2/2\sigma^2} P_X(x)\, dx. \tag{3}$$

Differentiating this with respect to y, and multiplying by σ^2 on both sides gives:

$$\begin{aligned} \sigma^2\, P'_Y(y) &= \int (x-y)\, \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-(y-x)^2/2\sigma^2}\, P_X(x)\, dx \\ &= \int (x-y)\, P_{X,Y}(x,y)\, dx \\ &= \int x\, P_{X,Y}(x,y)\, dx - y\, P_Y(y). \end{aligned}$$

Finally, dividing through by $P_Y(y)$, and combining with Eq. (1) gives

$$\begin{aligned} \hat{x}(y) &= y + \sigma^2 \frac{P'_Y(y)}{P_Y(y)} \\ &= y + \sigma^2 \frac{d}{dy} \log P_Y(y)\ . \end{aligned} \tag{4}$$

The important feature of the NEBLS estimator is its expression as a direct function of the measurement density, with no explicit reference to the prior. This means that the estimator can be approximated by estimating the measurement density from a set of observed samples. The derivation relies only on the assumptions of squared loss function, and additive Gaussian measurement noise, but is independent of the prior. We will assume squared loss throughout this article, but in Sec. 3, we describe a NEBLS form for more general measurement conditions.

We can gain some intuition for the solution by considering an example in which the prior distribution for x consists of three isolated point masses (delta functions). The measurement density may be obtained by convolving this prior with a Gaussian (top, Fig. 1). And, according to Eq. (4), the optimal estimator is obtained by adding the log derivative of the measurement density (bottom, Fig. 1) to the measurement. This is a form of gradient ascent, in which the estimator “shrinks” the observations toward the local maxima of the log density. In the vicinity of the most isolated (left) delta function, this shrinkage function is antisymmetric with a slope of negative one, resulting in essentially perfect recovery of the true value of x. Note that this optimal shrinkage is accomplished in a single step, unlike methods such as the mean-shift algorithm (Comaniciu and Meer, 2002), which uses iterative gradient ascent on the logarithm of a density to perform nonparametric clustering.

2.2 Dual formulation: SURE-optimized parametric least squares estimation

Next, as an alternative to the BLS estimator, consider the parametric regression formulation of the optimal estimation problem. Given a family of estimators, $f_{\boldsymbol{\theta}}$, parameterized by vector $\boldsymbol{\theta}$, we wish to select the one

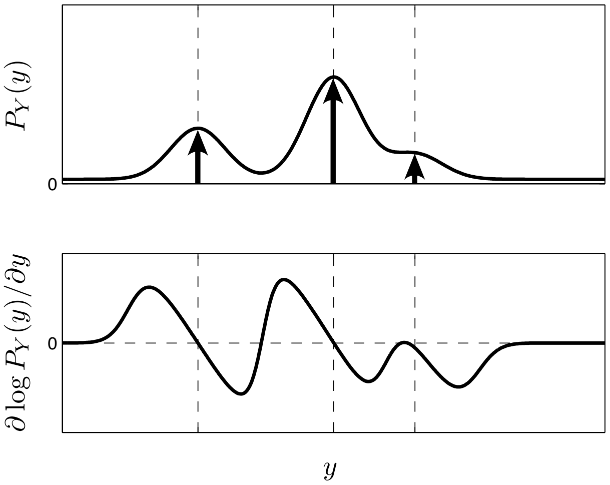


Fig. 1: Illustration of the NEBLS estimator for a one-dimensional value with additive Gaussian noise. **Top:** Measurement density, $P_Y(y)$, arising from a signal with prior consisting of three point masses (indicated by upward arrows) corrupted by additive Gaussian noise. **Bottom:** Derivative of the log measurement density, which (when added to the measurement), gives the NEBLS estimator (see Eq. (4)).

that minimizes the expected squared error:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \mathbf{E}_{X,Y}\left((f_{\boldsymbol{\theta}}(Y) - X)^2\right),$$

where the subscripts on the expectation indicate that it is taken over the joint density of measurements and correct values. In practice, the optimal parameters are obtained by approximating the expectation with a sum over clean/noisy pairs of data, $\{x_k, y_k\}$:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \frac{1}{N}\sum_{k=1}^{N}\Big(f_{\boldsymbol{\theta}}(y_k) - x_k\Big)^2. \tag{5}$$

Although this regression expression is written in terms of a supervised training set (i.e., one that includes the true values, x_k), the NEBLS formula of Eq. (4) may be used to derive an expression for the mean squared error that relies only on the noisy measurements, y_k. To see this, we rewrite the parametric estimator as $f_{\boldsymbol{\theta}}(y) = y + g_{\boldsymbol{\theta}}(y)$, and expand the mean squared error:

$$\begin{aligned}\mathbf{E}_{X,Y}\left((f_{\boldsymbol{\theta}}(Y) - X)^2\right) &= \mathbf{E}_{X,Y}\left((Y + g_{\boldsymbol{\theta}}(Y) - X)^2\right) \\ &= \mathbf{E}_{X,Y}\left(g^2_{\boldsymbol{\theta}}(Y)\right) + 2\mathbf{E}_{X,Y}\Big(g_{\boldsymbol{\theta}}(Y)\cdot(Y - X)\Big) + \mathbf{E}_{X,Y}\left((Y - X)^2\right) \\ &= \mathbf{E}_{Y}\left(g^2_{\boldsymbol{\theta}}(Y)\right) + 2\mathbf{E}_{X,Y}\Big(g_{\boldsymbol{\theta}}(Y)\cdot(Y - X)\Big) + \sigma^2. \end{aligned} \tag{6}$$

The middle term can be written as an expectation over Y alone, by substituting the NEBLS estimator from

Eq. (4), and then integrating by parts:

$$\begin{aligned}
\mathbf{E}_{X,Y}\left(g_{\boldsymbol{\theta}}(Y)\cdot(Y-X)\right) &= \mathbf{E}_{Y}\left(g_{\boldsymbol{\theta}}(Y)\cdot(Y-\mathbf{E}_{X|Y}\left(X|Y\right))\right)\\
&= \mathbf{E}_{Y}\left(g_{\boldsymbol{\theta}}(Y)\cdot(Y-Y-\sigma^2\frac{P_Y'(Y)}{P_Y(Y)})\right)\\
&= -\sigma^2\,\mathbf{E}_{Y}\left(g_{\boldsymbol{\theta}}(Y)\cdot\frac{P_Y'(Y)}{P_Y(Y)}\right)\\
&= -\sigma^2\int\left(g_{\boldsymbol{\theta}}(y)\cdot\frac{P_Y'(y)}{P_Y(y)}\right)P_Y(y)\,dy\\
&= -\sigma^2\int g_{\boldsymbol{\theta}}(y)\cdot P_Y'(y)\,dy\\
&= \sigma^2\int g_{\boldsymbol{\theta}}'(y)\cdot P_Y(y)\,dy\\
&= \sigma^2\,\mathbf{E}_{Y}\left(g_{\boldsymbol{\theta}}'(Y)\right),
\end{aligned}$$

where we have assumed that the term $P_Y(y)g_{\boldsymbol{\theta}}(y)|_{-\infty}^{\infty}$ that arises when integrating by parts is zero (i.e., as y goes to $\pm\infty$, we assume $P_Y(y)$ dies off faster than $g_{\boldsymbol{\theta}}(y)$ grows).

Finally, substituting this back into Eq. (6) gives an expression for estimation error:

$$\begin{aligned}
\mathbf{E}_{X,Y}\left(\left(X-f_{\boldsymbol{\theta}}(Y)\right)^2\right) &= \mathbf{E}_{Y}\left(g_{\boldsymbol{\theta}}^2(Y)\right)+2\sigma^2\mathbf{E}_{Y}\left(g_{\boldsymbol{\theta}}'(Y)\right)+\sigma^2\\
&= \mathbf{E}_{Y}\left(g_{\boldsymbol{\theta}}^2(Y)+2\sigma^2 g_{\boldsymbol{\theta}}'(Y)+\sigma^2\right). \qquad (7)
\end{aligned}$$

This remarkable equation expresses the mean squared error over the joint distribution of values and measurements as an expected value over the measurements alone. It is known as *Stein's unbiased risk estimator (SURE)*, after Charles Stein, who derived it[2] and used it as a means of comparing the quality of different estimators (Stein, 1981). In Sec. 4, we derive a much more general form of this expression that does not assume additive Gaussian noise.

In practice, SURE can be approximated as an average over an unsupervised set of noisy measurements, and this approximation can then be minimized over the parameters, $\boldsymbol{\theta}$, to select an estimator:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}}\frac{1}{N}\sum_{k=1}^{N}\left(g_{\boldsymbol{\theta}}^2(y_k)+2\sigma^2 g_{\boldsymbol{\theta}}'(y_k)\right), \qquad (8)$$

where we've dropped the last term (σ^2) because it does not depend on the parameter vector, $\boldsymbol{\theta}$. Note that unlike the supervised regression form of Eq. (5), this optimization does not rely on access to true signal values, $\{x_k\}$. We refer to the function associated with the optimized parameter, $g_{\hat{\boldsymbol{\theta}}}(y)$, as a *SURE-optimized parametric least squares* (*SURE2PLS*) estimator. This type of estimator was first developed for removal of additive Gaussian noise by Donoho and Johnstone (1995), who used it to estimate an optimal threshold

[2] Note that Stein derived the expression directly, without reference to Miyasawa's NEBLS estimator. In addition, Stein formulated the problem in a "frequentist" context where X is a fixed (nonrandom) but unknown parameter, and his result is expressed in terms of conditional expectations over $Y|X$. This may be readily obtained from our formulation by assuming a degenerate prior (Dirac delta) with mass at this fixed but unknown value, and replacing all expectations over $\{X, Y\}$ or Y with conditional expectations over $Y|X$. Conversely, Stein's formulation in terms of conditional densities may be easily converted into our result by taking expectations over X.

shrinkage function (the resulting estimator is named *SUREshrink*). These results have been generalized to other shrinkage estimators (Benazza-Benyahia and Pesquet, 2005; Chaux et al., 2008), as well as linear families of estimators (Pesquet and Leporini, 1997; Luisier et al., 2006; Raphan and Simoncelli, 2007b; Blu and Luisier, 2007; Raphan and Simoncelli, 2008), which we discuss in Sec. 5.

Intuitively, the objective function in Eq. (8) favors functions $g(\cdot)$ that have both a small squared magnitude and a large negative derivative in locations where the measurements, y_k, are concentrated. As such, a good estimator will "shrink" the data toward regions of high probability, as can be seen in the optimal solution shown in Fig. 1. Note also that the noise parameter σ^2 acts as a weighting factor on the derivative term, effectively controlling the smoothness of the solution.

2.3 Combining NEBLS with SURE yields the Score Matching density estimator

The SURE2PLS solution developed in the previous section can be applied to any parametric estimator. Consider a specific estimator that is derived by starting with a parametric form for the prior, $P_X^{(\phi)}(x)$. Combining this with the likelihood (using Eq. (3), which specifies a convolution of the Gaussian noise density with the prior) generates a parametric form for the measurement density, $P_Y^{(\phi)}(y)$. And given this, the expression of Eq. (4) gives the BLS estimator associated with this prior: $g_\phi(y) = \sigma^2 \frac{d}{dy} \log P_Y^{(\phi)}(y)$. Finally, substituting this estimator into Eq. (8) and eliminating common factors of σ yields an objective function for the density parameters ϕ given samples $\{y_k\}$:

$$\hat{\phi} = \arg\min_{\phi} \frac{1}{N} \sum_{k=1}^{N} \left[\left(\frac{d}{dy} \log P_Y^{(\phi)}(y_k) \right)^2 + 2 \frac{d^2}{dy^2} \log P_Y^{(\phi)}(y_k) \right]. \qquad (9)$$

This objective function allows one to choose density parameters that are optimal for solving the least-squares estimation problem. By comparison, most parametric empirical Bayes procedures select parameters for the prior density by optimizing some other criterion (e.g., maximizing likelihood of the data, or matching moments) (Casella, 1985), which are inconsistent with the estimation goal.[3]

Outside of the estimation context, Eq. (9) provides an objective function that can be used for estimating the parameters of the density $P_Y^{(\phi)}(y)$ from samples. This general methodology, dubbed *score matching*, was originally proposed by Hyvärinen (2005), who developed it by noting that differentiating the log of a density eliminates the normalization constant (known in physics as the "partition function"). This constant is generally a complicated function of the parameters, and thus an obstacle to solving the density estimation problem. In a subsequent publication (Hyvärinen, 2008), Hyvärinen showed a relationship of score matching to SURE, by assuming the density to be estimated is the prior (i.e., the density of the *clean* signal), and taking the limit as the variance of the Gaussian noise goes to zero. Here (and previously, in (Raphan and Simoncelli, 2007b)), we have interpreted score-matching as a means of estimating the density of the *noisy* measurements, with the objective function expressing the MSE achieved by an estimator that is optimal for removing *finite-variance* Gaussian noise when the measurements are drawn from $P_Y^{(\phi)}(y)$. Notice that in this context, although σ^2 does not appear in the objective function of Eq. (9), it provides a means of controlling the smoothness of the parametric density family, $P_Y^{(\phi)}(y)$, by Eq. (3). Of course, just as maximum likelihood (ML) has been used to estimate parametric densities outside the context in which it is optimal

[3] In particular, maximizing likelihood minimizes the Kullback-Leibler divergence between the true density and the parametric density (Wasserman, 2004).

(i.e., compression), the score matching methodology has been used in contexts for which squared estimation error is not relevant, and with parametric families that cannot have arisen from an additive Gaussian noise process.

3 General formulation: NEBLS estimator

We now develop a generalized form for the NEBLS estimator of Eq. (4). Suppose we make a vector observation, $\boldsymbol{y}$, that is a corrupted version of an unknown vector $\boldsymbol{x}$ (these need not have the same dimensionality). The BLS estimate is again the conditional expectation of the posterior density, which we can express using Bayes' rule as[4]

$$\begin{aligned} \hat{\boldsymbol{x}}(\boldsymbol{y}) &= \int \boldsymbol{x}\, P_{X|Y}(\boldsymbol{x}|\boldsymbol{y})\, d\boldsymbol{x} \\ &= \frac{\int \boldsymbol{x}\, P_{Y|X}(\boldsymbol{y}|\boldsymbol{x})\, P_X(\boldsymbol{x})\, d\boldsymbol{x}}{P_Y(\boldsymbol{y})}\,. \end{aligned} \tag{10}$$

Now define linear operator $\mathbf{A}$ to perform an inner product with the likelihood function

$$\mathbf{A}\{f\}(\boldsymbol{y}) \equiv \int P_{Y|X}(\boldsymbol{y}|\boldsymbol{x})\, f(\boldsymbol{x})\, d\boldsymbol{x},$$

and rewrite the measurement density in terms of this operator:

$$\begin{aligned} P_Y(\boldsymbol{y}) &= \int P_{Y|X}(\boldsymbol{y}|\boldsymbol{x})\, P_X(\boldsymbol{x})\, d\boldsymbol{x} \\ &= \mathbf{A}\{P_X\}(\boldsymbol{y}). \end{aligned} \tag{11}$$

Similarly, the numerator of the BLS estimator (Eq. (10)) may be rewritten as a composition of linear transformations applied to $P_X(\boldsymbol{x})$:

$$\begin{aligned} N(\boldsymbol{y}) &= \int P_{Y|X}(\boldsymbol{y}|\boldsymbol{x})\, \boldsymbol{x}\, P_X(\boldsymbol{x})\, d\boldsymbol{x} \\ &= (\mathbf{A} \circ \mathbf{X})\{P_X\}(\boldsymbol{y}), \end{aligned} \tag{12}$$

where operator $\mathbf{X}$ is defined as

$$\mathbf{X}\{f\}(\boldsymbol{x}) \equiv \boldsymbol{x} f(\boldsymbol{x}).$$

Note that Eq. (11) implies that P_Y always lies in the range of $\mathbf{A}$. Assuming for the moment that $\mathbf{A}$ is invertible, we can define $\mathbf{A}^{-1}$, an operator that inverts the observation process, recovering P_X from P_Y. The numerator can then be written as:

$$\begin{aligned} N(\boldsymbol{y}) &= (\mathbf{A} \circ \mathbf{X} \circ \mathbf{A}^{-1})\{P_Y\}(\boldsymbol{y}) \\ &= \mathbf{L}\{P_Y\}(\boldsymbol{y}). \end{aligned} \tag{13}$$

with linear operator $\mathbf{L} \equiv \mathbf{A} \circ \mathbf{X} \circ \mathbf{A}^{-1}$. In general, this operator maps a scalar-valued function of a vector, $P_Y(\boldsymbol{y})$, to a vector-valued function. In the discrete scalar case, $P_Y(\boldsymbol{y})$ and $N(\boldsymbol{y})$ are each vectors, the

[4] The derivations throughout this article are written assuming continuous variables, but they hold for discrete variables as well, for which the integrals must be replaced by sums, functions by vectors, and functionals by matrices.

argument $\boldsymbol{y}$ selects a particular index, $\mathbf{A}$ is a matrix containing $\mathbf{P}_{Y|X}$, $\mathbf{X}$ is a diagonal matrix containing values of $\mathbf{x}$, and $\circ$ is simply matrix multiplication. Combining all of these, we arrive at a general NEBLS form of the estimator:

$$\hat{\boldsymbol{x}}(\boldsymbol{y}) = \frac{\mathbf{L}\{P_Y\}(\boldsymbol{y})}{P_Y(\boldsymbol{y})}. \tag{14}$$

That is, the BLS estimator may be computed by applying a linear operator to the measurement density, and dividing this by the measurement density. This linear operator is determined solely by the observation process (as specified by the density $P_{Y|X}$), and thus the estimator does not require any knowledge of or assumption about the prior P_X.

The derivation given above may seem like sleight of hand, given that we assumed that the prior could be recovered exactly from the measurement density using $\mathbf{A}^{-1}$. But in Sec. 6.2, we show that while the operator $\mathbf{A}^{-1}$ may be ill-conditioned (e.g., a deconvolution in our introductory example of Sec. 2), it is often the case that the composite operator $\mathbf{L}$ is more stable (e.g., a derivative in the introductory example). More surprisingly, even when $\mathbf{A}$ is strictly non-invertible, it may still be possible to find an operator $\mathbf{L}$ that generates a NEBLS estimator. In Sec. 6 and Appendix B, we use Eq. (14) to derive NEBLS estimators for specific observation models, including those that have appeared in previous literature.

The NEBLS expression of Eq. (14) may be used to rewrite other expectations over X in terms of expectations over Y. For example, if we wish to calculate $E_{X|Y}\{X^n|Y = \boldsymbol{y}\}$, then Eq. (13) would be replaced by $(\mathbf{A} \circ \mathbf{X}^n \circ \mathbf{A}^{-1})\{P_Y\} = (\mathbf{A} \circ \mathbf{X} \circ \mathbf{A}^{-1})^n\{P_Y\} = \mathbf{L}^n\{P_Y\}$.[5] Exploiting the linearity of the conditional expectation, we may extend this to *any* polynomial function, $g(x) = \sum c_k x^k$:

$$\begin{aligned}
\mathbf{E}_{X|Y}\left(g(X)|Y = \boldsymbol{y}\right) &\approx \mathbf{E}_{X|Y}\left(\sum_k c_k X^k \middle| Y = \boldsymbol{y}\right) \\
&= \frac{\sum_k c_k \mathbf{L}^k\{P_Y\}(\boldsymbol{y})}{P_Y(\boldsymbol{y})} \\
&= \frac{g(\mathbf{L})\{P_Y\}(\boldsymbol{y})}{P_Y(\boldsymbol{y})}. \qquad (15)
\end{aligned}$$

And finally, consider the problem of finding the BLS estimator of X given Z where

$$Z = r(Y),$$

with r an invertible, differentiable, transformation. Using the known properties of change of variables for densities, we can write $P_Y(Y) = J_r P_Z(r(Y))$, where J_r is the Jacobian of the transformation $r(\cdot)$. From this, we obtain

$$\begin{aligned}
\mathbf{E}_{X|Z}\left(X|Z = \boldsymbol{z}\right) &= \mathbf{E}_{X|Y}\left(X|Y = r^{-1}(\boldsymbol{z})\right) \\
&= \frac{\mathbf{L}\{P_Y\}(r^{-1}(\boldsymbol{z}))}{P_Y(r^{-1}(\boldsymbol{z}))} \\
&= \frac{\mathbf{L}\left\{J_r(P_Z \circ r)\right\}(r^{-1}(\boldsymbol{z}))}{J_r P_Z(\boldsymbol{z})}. \qquad (16)
\end{aligned}$$

[5] It is easiest to imagine this in the scalar case, but the formula is also valid for the vector case, where n becomes a multi-index.

4 Dual formulation: SURE2PLS estimation

As in the scalar Gaussian case, the NEBLS estimator may be used to develop an expression for the mean squared error that does not depend explicitly on the prior, and this may be used to select an optimal estimator from a parametric family. This form is particularly useful in cases where it proves difficult to develop a stable nonparametric approximation of the ratio in Eq. (14) (Carlin and Louis, 2009).

Consider an estimator $\boldsymbol{f_\theta}(Y)$ parameterized by vector $\boldsymbol{\theta}$, and expand the mean squared error as:

$$\mathbf{E}_{X,Y}\left(|\boldsymbol{f_\theta}(Y) - X|^2\right) \;=\; \mathbf{E}_{X,Y}\left(|\boldsymbol{f_\theta}(Y)|^2 - 2\boldsymbol{f_\theta}(Y)\cdot X + |X|^2\right). \tag{17}$$

Using the NEBLS estimator of Eq. (14), the second term of the expectation may be written as

$$\begin{aligned}
\mathbf{E}_{X,Y}\left(\boldsymbol{f_\theta}(Y)\cdot X\right) &= \mathbf{E}_Y\left(\boldsymbol{f_\theta}(Y)\cdot \mathbf{E}_{X|Y}\left(X|Y\right)\right) \\
&= \mathbf{E}_Y\left(\boldsymbol{f_\theta}(Y)\cdot \frac{\mathbf{L}\{P_Y\}(Y)}{P_Y(Y)}\right) \\
&= \int \boldsymbol{f_\theta}(\boldsymbol{y})\cdot \frac{\mathbf{L}\{P_Y\}(\boldsymbol{y})}{P_Y(\boldsymbol{y})} P_Y(\boldsymbol{y})\, d\boldsymbol{y} \\
&= \int \boldsymbol{f_\theta}(\boldsymbol{y})\cdot \mathbf{L}\{P_Y\}(\boldsymbol{y})\, d\boldsymbol{y} &\text{(18a)} \\
&= \int \mathbf{L}^*\{\boldsymbol{f_\theta}\}(\boldsymbol{y})\cdot P_Y(\boldsymbol{y})\, d\boldsymbol{y} &\text{(18b)} \\
&= \mathbf{E}_Y\left(\mathbf{L}^*\{\boldsymbol{f_\theta}\}(Y)\right), &\text{(18c)}
\end{aligned}$$

where $\mathbf{L}^*$ is the dual operator of $\mathbf{L}$, defined by the equality of lines (18a) and (18b). In general, $\mathbf{L}^*$ maps a vector-valued function of $\boldsymbol{y}$ into a scalar-valued function of $\boldsymbol{y}$. In the discrete scalar case, $\mathbf{L}^*$ is simply the matrix transpose of $\mathbf{L}$. Substituting this for the second term of Eq. (17), and dropping the last term (since it does not depend on $\boldsymbol{\theta}$), gives a prior-free expression for the optimal parameter vector:

$$\begin{aligned}
\hat{\boldsymbol{\theta}} &= \arg\min_{\boldsymbol{\theta}} \mathbf{E}_{X,Y}\left(|\boldsymbol{f_\theta}(Y) - X|^2\right) \\
&= \arg\min_{\boldsymbol{\theta}} \mathbf{E}_Y\left(|\boldsymbol{f_\theta}(Y)|^2 - 2\mathbf{L}^*\{\boldsymbol{f_\theta}\}(Y)\right). &\text{(19)}
\end{aligned}$$

This *generalized SURE* (gSURE) form includes as special cases those formulations that have appeared in previous literature[6] (see introduction, and Table 1, for specific citations). Our approach thus serves to unify and generalize these results, and to show that they can be derived from the corresponding NEBLS estimators. Conversely, it is relatively straightforward to show that the estimator of Eq. (14) can be derived from the gSURE expression of Eq. (19) (see (Raphan, 2007) for a proof).

In practice, we can solve for the optimal $\boldsymbol{\theta}$ by minimizing the sample mean of this quantity:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \frac{1}{N}\sum_{k=1}^{N}\left\{|\boldsymbol{f_\theta}(\boldsymbol{y}_k)|^2 - 2\mathbf{L}^*\{\boldsymbol{f_\theta}\}(\boldsymbol{y}_k)\right\}. \tag{20}$$

[6] As in the Gaussian (SURE) case, these previous results were described in a frequentist setting in which X is fixed but unknown, and are written as expectations over $Y|X$, but they are equivalent to our results (see footnote 2). In practice, there is no difference between assuming the clean data are i.i.d. samples from a prior distribution, or assuming they are fixed and unknown values that are to be estimated individually from their corresponding Y values.

where $\{\boldsymbol{y}_k\}$ is a set of observed data. This optimization does not require any knowledge of (or samples drawn from) the prior P_X, and so we think of it as the unsupervised counterpart of the standard (supervised) regression solution of Eq. (5). When trained on insufficient data, this estimator can still exhibit errors analogous to overfitting errors seen in supervised training. This is because the sample mean in Eq. (20) is only asymptotically equal to the MSE. As with supervised regression, cross-validation or other resampling methods can be used to limit the dimensionality or complexity of the parameterization so that it is appropriate for the available data.

The resulting *generalized SURE-optimized parametric least squares* (gSURE2PLS) estimator, $\boldsymbol{f}_{\hat{\boldsymbol{\theta}}}$, may be applied to the same data set that is used to optimize $\hat{\boldsymbol{\theta}}$. This may seem odd when compared to supervised regression, for which such a statement makes no sense (since the supervised training set already includes the correct answers). But in the unsupervised context, each newly acquired measurement can be used for both estimation *and* learning, and it would be wasteful not to take advantage of this fact. In cases where one also has access to some supervised data $\{\boldsymbol{x}_n, \boldsymbol{y}_n | n \in \mathcal{S}\}$, in addition to unsupervised data $\{\boldsymbol{y}_n | n \in \mathcal{U}\}$, the corresponding objective functions may be combined additively (since they both represent squared errors) to obtain a semi-supervised solution:

$$\begin{aligned}
\hat{\boldsymbol{\theta}} &= \arg\min_{\boldsymbol{\theta}} \sum_{k\in\mathcal{S}} |\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{y}_k) - \boldsymbol{x}_k|^2 + \sum_{k\in\mathcal{U}} \Big\{ |\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{y}_k)|^2 - 2\mathbf{L}^*\{\boldsymbol{f}_{\boldsymbol{\theta}}\}(\boldsymbol{y}_k) \Big\} \\
&= \arg\min_{\boldsymbol{\theta}} \sum_{k\in\mathcal{S}\cup\mathcal{U}} |\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{y}_k)|^2 - 2\sum_{k\in\mathcal{S}} \boldsymbol{x}_k \boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{y}_k) - 2\sum_{k\in\mathcal{U}} \mathbf{L}^*\{\boldsymbol{f}_{\boldsymbol{\theta}}\}(\boldsymbol{y}_k),
\end{aligned}$$

where we've again discarded a term that does not depend on $\boldsymbol{\theta}$. Again, the gSURE2PLS methodology allows the estimator to be initialized with some supervised training data, but then to continue to adapt its estimator *while* performing the estimation task.

The operator $\mathbf{L}^*$ extracts that information from a function on Y that is relevant for estimating X, and may be used in more general settings than the one considered here. For example, the derivation in Eq. (18c) can be generalized, using the result in Eq. (15), to give

$$\mathbf{E}_{X,Y}\left(f(X)g(Y)\right) \quad = \quad \mathbf{E}_Y\left(f(\mathbf{L}^*)\{g\}(Y)\right),$$

for arbitrary polynomials f and g (and hence functions that are well-approximated by polynomials). And this may be further generalized to any joint polynomial, which can be written as a sum of pairwise products of polynomials in each variable. As a particular example, this means that we can recover any statistic of X (e.g., any moment of the prior) through an expectation over Y:

$$\mathbf{E}_X\left(f(X)\right) = \mathbf{E}_Y\left(f(\mathbf{L}^*)\{\mathbf{1}\}(Y)\right), \tag{21}$$

where $\mathbf{1}$ indicates a function whose value is (or vector whose components are) always one.

Finally, for a parametric prior density, the NEBLS estimator of Eq. (14) may be substituted into the gSURE objective function of Eq. (20) to obtain a generalized form of the score matching density estimator of Eq. (9). Specifically, a parametric density $P_Y^{(\phi)}$ may be fit to data $\{\boldsymbol{y}_k\}$ by solving

$$\hat{\phi} = \arg\min_{\phi} \frac{1}{N}\sum_{k=1}^{N}\left[\left|\frac{\mathbf{L}\{P_Y^{(\phi)}\}}{P_Y^{(\phi)}}(\boldsymbol{y}_k)\right|^2 - 2\mathbf{L}^*\left\{\frac{\mathbf{L}\{P_Y^{(\phi)}\}}{P_Y^{(\phi)}}\right\}(\boldsymbol{y}_k)\right]. \tag{22}$$

Recall that the operator $\mathbf{L}$ is determined by the observation process that governs the relationship between the true signal and the measurements in the original estimation problem. When used in this parametric density estimation context, different choices of operator will lead to different density estimators, in which the density is selected from a family "smoothed" by the measurement process underlying $\mathbf{L}$. In all cases, the density estimation problem can be solved without computing the normalization factor, which is eliminated in the quotient $\mathbf{L}\{P_Y^{(\phi)}\}/P_Y^{(\phi)}$.

As a specific example, assume the observations are positive integers, n, sampled from a mixture of Poisson densities, $P_Y^{(\phi)}(n)$, where the rate variable X is distributed according to a parametric prior density, $P_X^{(\phi)}(x)$. Using the form of $\mathbf{L}$ for Poisson observations (see Sec. 6.1), we obtain an estimator for parameter ϕ from data $\{n_k\}$:

$$\hat{\phi} = \arg\min_{\phi} \frac{1}{N} \sum_k \left\{ \left(\frac{(n_k+1)P_Y^{(\phi)}(n_k+1)}{P_Y^{(\phi)}(n_k)} \right)^2 - 2\left(\frac{(n_k)^2 P_Y^{(\phi)}(n_k)}{P_Y^{(\phi)}(n_k-1)} \right) \right\}.$$

5 Incremental gSURE2PLS optimization for kernel estimators

The gSURE2PLS methodology introduced in Sec. 4 requires us to minimize an expression that is quadratic in the estimation function. This makes it particularly appealing for use with estimators that are *linear* in their parameters, and several authors have exploited this in developing estimators for the additive Gaussian noise case (Pesquet and Leporini, 1997; Luisier et al., 2006; Raphan and Simoncelli, 2007b; Blu and Luisier, 2007; Raphan and Simoncelli, 2008). Here, we show that this advantage holds for the general case, and we use it to develop an incremental algorithm for optimizing the estimator.

Consider a scalar estimator that is formed as a weighted sum of fixed nonlinear kernel functions:

$$f_{\boldsymbol{\theta}}(y) = \sum_j \theta_j \; h_j(y) = \boldsymbol{\theta}^T \boldsymbol{h}(y),$$

where $\boldsymbol{h}(y)$ is a vector with components containing the kernel functions, $h_j(y)$. Substituting this into Eq. (20), and using the linearity of the operator $\mathbf{L}^*$ gives a gSURE expression for unsupervised parameter optimization from n samples:

$$\hat{\boldsymbol{\theta}}_n = \arg\min_{\boldsymbol{\theta}} \frac{1}{n} \left\{ \boldsymbol{\theta}^T \left(\sum_{k=1}^{n} \boldsymbol{h}(y_k)\boldsymbol{h}(y_k)^T \right) \boldsymbol{\theta} - 2\boldsymbol{\theta}^T \sum_{k=1}^{n} \mathbf{L}^*\{\boldsymbol{h}\}(y_k) \right\}, \tag{23}$$

where $\mathbf{L}^*\{\boldsymbol{h}\}(y)$ is a vector whose j^{th} component is $\mathbf{L}^*\{h_j\}(y)$. The quadratic form of the objective function allows us to write the solution as a familiar closed-form expression:

$$\hat{\boldsymbol{\theta}}_n = \mathbf{C}_n^{-1} \boldsymbol{m}_n, \tag{24}$$

where we define

$$\mathbf{C}_n \equiv \sum_{k=1}^{n} \boldsymbol{h}(y_k)\boldsymbol{h}(y_k)^T \tag{25a}$$

$$\boldsymbol{m}_n \equiv \sum_{k=1}^{n} \mathbf{L}^*\{\boldsymbol{h}\}(y_k). \tag{25b}$$

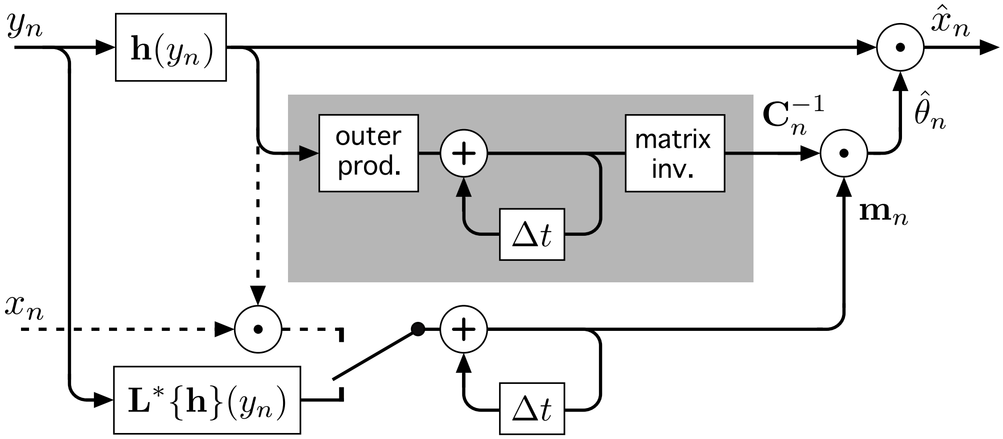


Fig. 2: Circuit diagram for an incremental gSURE2PLS kernel estimator, as specified by Eqs. (24) and (26). The quantities $\boldsymbol{m}_n$ and $\mathbf{C}_n$ must be accumulated and stored internally. The gray box encloses the portion of the diagram in which $\mathbf{C}_n^{-1}$ is computed, and may be replaced by a circuit that directly accumulates the inverse matrix, $\mathbf{S}_n = \mathbf{C}_n^{-1}$ (see Appendix A). The entire diagram can also be formulated in terms of the parameter vector, $\hat{\boldsymbol{\theta}}_n$, in place of the vector $\boldsymbol{m}_n$ (also shown in Appendix A). If supervised data (i.e., clean values x_n) are available for some n, they may beincorporated by flipping the switch to activate the dashed-line portion of the circuit (effectively replacing the right hand side of Eq. (18c) by the left hand side).

Note that the quantity $\boldsymbol{m}_n$ is (n times) an n-sample estimate of $\mathbf{E}_Y\left(\mathbf{L}^*\{\boldsymbol{h}\}(Y)\right)$, which by Eq. (18c) provides an unsupervised estimate of $\mathbf{E}_{X,Y}\left(\boldsymbol{h}(Y) \cdot X\right)$.

In addition to allowing a direct solution, the quadratic form of this objective function lends itself to an incremental algorithm, in which the estimator is both applied to and updated by each measurement as it is acquired sequentially over time. The advantage of such a formulation is that the estimator can be updated gradually, based on all previous observations, but without needing to store and access those observations for each update. To see this, we rewrite the expressions in Eq. (25) as:

$$\mathbf{C}_n = \mathbf{C}_{n-1} + \boldsymbol{h}(y_n)\,\boldsymbol{h}(y_n)^T \tag{26a}$$

$$\boldsymbol{m}_n = \boldsymbol{m}_{n-1} + \mathbf{L}^*\{\boldsymbol{h}\}(y_n). \tag{26b}$$

These equations, combined with Eq. (24), imply that the optimal parameter vector can be computed by combining the most recent data observation $\boldsymbol{y}_n$ with summary information stored in an accumulated matrix $\mathbf{C}_{n-1}$ and vector $\boldsymbol{m}_{n-1}$. The basic algorithm is illustrated in a flow diagram in Fig. 2. For each iteration of the algorithm, an incoming data value y_n is used to incrementally update the estimator, and the same estimator is then used to compute the estimate $\hat{x}_n$. The diagram also includes an optional path for supervised data x_n, which may be used to improve the estimator.

This incremental formulation bears some similarity to the well-known Kalman filter (Kalman, 1960), which provides an incremental estimate of a state variable that is observed through additive Gaussian measurements. But note that the standard Kalman filter is based on a state model with known linear/Gaussian dynamics and known Gaussian noise observations, whereas our formulation allows a more general (although still assumed known) observation model that is applied to independent state values drawn independently from an unknown prior.

In practice, this algorithm may be made more efficient and flexible in a number of ways. Specifically, the equations may be re-written so as to accumulate the inverse covariance matrix, thus avoiding the costly matrix inversion on each iteration. They may also be written so as to directly accumulate the estimator parameter vector $\hat{\boldsymbol{\theta}}$, instead of $\boldsymbol{m}$. Finally, the accumulation of data can be weighted (e.g., exponentially) over time, so as to emphasize the most recent data and "forget" older data. This is particularly useful in a case where the distribution of the state variable is changing over time. These variations are described in Appendix A.

6 Derivation of NEBLS estimators

In Sec. 2, we developed the NEBLS estimator for the scalar case of additive Gaussian noise. The formulation of Sec. 3 is much more general, and can be used to obtain NEBLS estimators for a variety of other corruption processes. But in many cases, it is difficult to obtain the operator $\mathbf{L}$ directly from the expression in Eq. (13), because inversion of the operator $\mathbf{A}$ is unstable or undefined. This issue is addressed in detail in Sec. 6.2. But we first note that it is necessary and sufficient that applying $\mathbf{L}$ to the measurement density produce the numerator of the BLS estimator in Eq. (10):

$$(\mathbf{L} \circ \mathbf{A})\{P_X\} = (\mathbf{A} \circ \mathbf{X})\{P_X\},$$

where we've used Eqs. (11) and (12) to express both numerator and measurement density as linear functions of the prior. This expression must hold for every prior density, P_X. From the definition of $\mathbf{A}$, this means that it is sufficient to find an operator $\mathbf{L}$ such that

$$\mathbf{L}\{P_{Y|X}(\boldsymbol{y}|\boldsymbol{x})\} = \boldsymbol{x} P_{Y|X}(\boldsymbol{y}|\boldsymbol{x}). \tag{27}$$

Cressie (1982) noted previously that if an operator could be found that satisfied this relationship, then it could be used to define a corresponding NEBLS estimator. Here, we note that in the scalar case, this equation implies that for each value of $\boldsymbol{x}$, the conditional density $P_{Y|X}(\boldsymbol{y}|\boldsymbol{x})$ must be an *eigenfunction* (or eigenvector, for discrete variables) of operator $\mathbf{L}$, with associated eigenvalue $\boldsymbol{x}$.[7] We have used this eigenfunction property to obtain a variety of NEBLS estimators by direct inspection of the observation density $P_{Y|X}$. Table 1 provides a listing of these, including those examples that appear previously in the nonparametric empirical Bayes and SURE-optimized estimator literatures.

6.1 Derivation of specific NEBLS estimators

In this section, we provide a derivation for a few specific NEBLS estimators and their gSURE2PLS counterparts. Derivations of the remainder of the estimators given in Table 1 are provided in appendix B.

Additive noise: general case. Consider the case in which a vector-valued variable of interest is corrupted by independent additive noise: $Y = X + W$, with the noise drawn from some distribution, $P_W(\boldsymbol{w})$. The conditional density may then be written

$$P_{Y|X}(\boldsymbol{y}|\boldsymbol{x}) = P_W(\boldsymbol{y} - \boldsymbol{x}).$$

[7] For the vector case, this must be true for each component of $\boldsymbol{x}$ and associated component of the operator.

Obs. process		Obs. density: $P_{Y\mid X}(\boldsymbol{y}\mid\boldsymbol{x})$	Numerator of estimator: $\mathbf{L}\{P_Y\}(\boldsymbol{y})$
General discrete		$\mathbf{A}$ (matrix)	$(\mathbf{A} \circ X \circ \mathbf{A}^{-1})P_Y(\boldsymbol{y})$
Additive	General (6.1)	$P_W(\boldsymbol{y}-\boldsymbol{x})$	$\boldsymbol{y}P_Y(\boldsymbol{y})$ $-\mathcal{F}^{-1}\left\{i\nabla_{\boldsymbol{\omega}} \ln\left(\widehat{P_W}(\boldsymbol{\omega})\right)\widehat{P_Y}(\boldsymbol{\omega})\right\}(\boldsymbol{y})$
	Gaussian [Miyasawa61, Stein81*]	$\frac{\exp -\frac{1}{2}(\mathbf{y}-\mathbf{x}-\mu)^T\Lambda^{-1}(\mathbf{y}-\mathbf{x}-\mu)}{\sqrt{\lvert 2\pi\Lambda\rvert}}$	$(\boldsymbol{y}-\boldsymbol{\mu})P_Y(\boldsymbol{y}) + \Lambda\nabla_{\boldsymbol{y}}P_Y(\boldsymbol{y})$
	Laplacian	$\frac{1}{2\alpha}e^{-\lvert(y-x)/\alpha\rvert}$	$yP_Y(y) + 2\alpha^2\{P'_W \star P_Y\}(y)$
	Poisson	$\sum \frac{\lambda^k e^{-\lambda}}{k!}\delta(y-x-ks)$	$yP_Y(y) - \lambda s P_Y(y-s)$
	Cauchy	$\frac{1}{\pi}(\frac{\alpha}{(\alpha(y-x))^2+1})$	$yP_Y(y) - \{\frac{1}{2\pi\alpha y} \star P_Y\}(y)$
	Uniform	$\begin{cases} \frac{1}{2a}, & \lvert y-x\rvert \leq a \\ 0, & \lvert y-x\rvert > a \end{cases}$	$yP_Y(y) + a\sum_k \text{sgn}(k)P_Y(y-ak)$ $-\ \frac{1}{2}\int P_Y(\tilde{y})\text{sgn}(y-\tilde{y})d\tilde{y}$
	Random # components	$P_W(y-x)$, where: $W \sim \sum_{k=0}^{K} W_k$, W_k i.i.d. (P_c), $\quad K \sim Poiss(\lambda)$	$yP_Y(y) - \lambda\{(yP_c) \star P_Y\}(y)$
	Gaussian scale mixture	$\boldsymbol{Y} = \boldsymbol{x} + \sqrt{Z}\boldsymbol{U}$, $\boldsymbol{U} \sim N(0,\Lambda), \quad Z \sim p_Z$	$\boldsymbol{y}P_Y(\boldsymbol{y})+$ $\left(\mathcal{F}^{-1}\left\{\frac{\int_0^\infty zp_z(z)e^{-z\frac{1}{2}\boldsymbol{\omega}^T\Lambda\boldsymbol{\omega}}dz}{\int_0^\infty p_z(z)e^{-z\frac{1}{2}\boldsymbol{\omega}^T\Lambda\boldsymbol{\omega}}dz}\right\} \star \Lambda\nabla_{\boldsymbol{y}}P_Y\right)(\boldsymbol{y})$
Disc. Exp.	General (B.1) [Maritz89, Hwang82*]	$h(x)g(n)x^n$	$\frac{g(n)}{g(n+1)}P_Y(n+1)$
	Inverse [Hwang82*]	$h(x)g(n)x^{-n}$	$\frac{g(n)}{g(n-1)}P_Y(n-1)$
	Poisson (6.1) [Robbins56, Hwang82*]	$\frac{x^n e^{-x}}{n!}$	$(n+1)P_Y(n+1)$
Cts. Exp.	General (B.2) [Maritz89, Berger80*]	$h(x)g(y)e^{T(y)x}$	$\frac{g(y)}{T'(y)}\frac{d}{dy}\left(\frac{P_Y(y)}{g(y)}\right)$
	Inverse [Berger80*]	$h(x)g(y)e^{T(y)/x}$	$-g(y)\int_y^\infty \frac{T'(\tilde{y})}{g(\tilde{y})}P_Y(\tilde{y})d\tilde{y}$
	Laplacian scale mixture	$\frac{1}{x}e^{-\frac{y}{x}}, \qquad x, y > 0$	$Pr\{Y > y\}$
Power of fixed	General (B.3)	$\widehat{P_{Y\mid X}}(\omega\mid x) = [\widehat{P_W}(\omega)]^x$	$\left(\mathcal{F}^{-1}\left\{\frac{1}{i\frac{d}{d\omega}\ln(\widehat{P_W}(\omega))}\right\} \star (yP_Y)\right)(y)$
	Gaussian scale mixture	$\frac{1}{\sqrt{2\pi x}}e^{-\frac{y^2}{2x}}$	$E_Y\{Y; Y > y\}$
	Signal-dependent AWGN	$Y = ax + \sqrt{x}W, \quad W \sim \mathcal{N}(0,1)$	$\text{sgn}(a)\left((e^{ax}\ I_{\{ax<0\}}) \star (yP_Y)\right)(y)$
	Multiplicative α-stable	$Y = x^{\frac{1}{\alpha}}W, \quad W$ α-stable	$\left(\mathcal{F}^{-1}\left\{\frac{i}{\text{sgn}(\omega)\lvert\omega\rvert^{\alpha-1}}\right\} \star (yP_Y)\right)(y)$
Multiplicative lognormal (B.4)		$Y = xe^W, \quad W$ Gaussian	$e^{\frac{3}{2}\sigma^2}P_Y(e^{\sigma^2}y)y$
Uniform mixture (B.5)		$\frac{1}{2x}, \quad \lvert y\rvert \leq x$ $0, \quad \lvert y\rvert > x$	$\lvert y\rvert P_Y(y) + Pr\{Y > \lvert y\rvert\}$

Table 1: NEBLS estimation formulas for a variety of observation processes, as listed in the left column. Parenthesized expressions indicate the section containing the derivation, and brackets contain bibliographic references for operators $\mathbf{L}$, with * denoting references for the parametric (dual) operator, $\mathbf{L}^*$. Middle column gives the measurement density (note that variable n replaces y for discrete measurements). Right column gives the numerator of the NEBLS estimator, $\hat{x}(\boldsymbol{y}) = \frac{\mathbf{L}\{P_Y\}(\boldsymbol{y})}{P_Y(\boldsymbol{y})}$. The symbol $\star$ indicates convolution, a hat (e.g., $\widehat{P_W}$) indicates a Fourier transform, and $\mathcal{F}^{-1}$ the inverse Fourier transform.

We seek an operator which, when applied to this conditional density (viewed as a function of $\boldsymbol{y}$), will obey

$$\mathbf{L}\{P_W(\boldsymbol{y}-\boldsymbol{x})\} = \boldsymbol{x}\, P_W(\boldsymbol{y}-\boldsymbol{x}). \tag{28}$$

Subtracting $\boldsymbol{y}\, P_W(\boldsymbol{y}-\boldsymbol{x})$ from both sides of Eq. (28) gives:

$$\mathbf{M}\left\{P_W(\boldsymbol{y}-\boldsymbol{x})\right\} = -(\boldsymbol{y}-\boldsymbol{x})\, P_W(\boldsymbol{y}-\boldsymbol{x}).$$

where we've defined linear operator $\mathbf{M}\left\{f\right\}(\boldsymbol{y}) \equiv \mathbf{L}\{P_Y\}(\boldsymbol{y}) - \boldsymbol{y}\, P_Y(\boldsymbol{y})$. Since this equation must hold for all $\boldsymbol{x}$, it implies that $\mathbf{M}$ is a linear *shift-invariant* operator (acting in $\boldsymbol{y}$), and can be represented as convolution with a kernel $\boldsymbol{m}(\boldsymbol{y})$. Taking the Fourier transform of both sides, and using the convolution and differentiation properties gives:

$$\begin{aligned} \widehat{\boldsymbol{m}}(\boldsymbol{\omega})\widehat{P_W}(\boldsymbol{\omega}) &= -\widehat{(\boldsymbol{y}P_W)}(\boldsymbol{\omega}) \\ &= -i\nabla_{\boldsymbol{\omega}}\widehat{P_W}(\boldsymbol{\omega}), \end{aligned}$$

so that

$$\begin{aligned} \widehat{\boldsymbol{m}}(\boldsymbol{\omega}) &= -i\frac{\nabla_{\boldsymbol{\omega}}\widehat{P_W}(\boldsymbol{\omega})}{\widehat{P_W}(\boldsymbol{\omega})} \\ &= -i\nabla_{\boldsymbol{\omega}} \ln\left(\widehat{P_W}(\boldsymbol{\omega})\right). \end{aligned} \tag{29}$$

Thus, the linear operator is:

$$\mathbf{L}\{P_Y\}(\boldsymbol{y}) = \boldsymbol{y}\, P_Y(\boldsymbol{y}) + \mathcal{F}^{-1}\left\{\widehat{\boldsymbol{m}}(\boldsymbol{\omega})\, \widehat{P_Y}(\boldsymbol{\omega})\right\}(\boldsymbol{y}), \tag{30}$$

where $\mathcal{F}^{-1}$ denotes the inverse Fourier transform. The dual operator (consistent with Eqs. (18a) and (18b)), is then:

$$\mathbf{L}^*\{\boldsymbol{f}\}(\boldsymbol{y}) = \boldsymbol{y}^T \boldsymbol{f}(\boldsymbol{y}) + \mathcal{F}^{-1}\left\{\widetilde{\widehat{\boldsymbol{m}}}(\boldsymbol{\omega})\, \widehat{\boldsymbol{f}}(\boldsymbol{\omega})\right\}(\boldsymbol{y}), \tag{31}$$

where $\boldsymbol{f}$ is the vector-valued estimator, and tilde indicates the Hermitian transpose (i.e. conjugation and transpose).

Note that throughout this discussion X and W play symmetric roles. Thus, we can solve for the BLS estimator if we know the density of *either* the noise, or the signal. We also note that if the additive noise is such that the observation process is not invertible (e.g., if the Fourier transform of P_W is bandlimited), the proof of Eq. (30) implies that this operator is still valid if we define

$$\nabla_{\boldsymbol{\omega}} \ln\left(\widehat{P_W}(\boldsymbol{\omega})\right) = 0,$$

whenever $\widehat{P_W}(\boldsymbol{\omega}) = 0$. As examples, we consider the following specific cases of additive noise.

Additive Gaussian noise (vector case). The expression in Eq. (30) can be used to derive the solution for the full vector-valued generalization of the scalar Gaussian additive noise case given in Sec. 2. The noise model is:

$$P_W(\boldsymbol{x}) = \frac{1}{(2\pi|\Lambda|)^{n/2}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Lambda^{-1}(\boldsymbol{x}-\boldsymbol{\mu})},$$

with covariance matrix Λ, and mean vector $\boldsymbol{\mu}$. In this case, the Fourier transform of the density is:

$$\widehat{P_W}(\boldsymbol{\omega}) \propto e^{-i\boldsymbol{\omega}^T\boldsymbol{\mu}-\frac{1}{2}\boldsymbol{\omega}^T\Lambda\boldsymbol{\omega}},$$

which, upon substitution into Eq. (29) yields:

$$\widehat{m}(\boldsymbol{\omega}) = \left[i\Lambda\boldsymbol{\omega} - \boldsymbol{\mu}\right].$$

Substituting into Eq. (30), and then into Eq. (14) yields the NEBLS estimator:

$$\begin{aligned} \mathbf{E}_{X|Y}\left(\boldsymbol{x}|Y=\boldsymbol{y}\right) &= \boldsymbol{y} - \boldsymbol{\mu} + \frac{\Lambda\nabla_{\boldsymbol{y}}P_Y(\boldsymbol{y})}{P_Y(\boldsymbol{y})} \\ &= \boldsymbol{y} - \boldsymbol{\mu} + \Lambda\nabla_{\boldsymbol{y}}\ln(P_Y(\boldsymbol{y})). \end{aligned} \tag{32}$$

And substituting into Eq. (31) yields the dual operator:

$$\mathbf{L}^*\{\boldsymbol{f}\}(\boldsymbol{y}) = (\boldsymbol{y}-\boldsymbol{\mu})^T\boldsymbol{f}(\boldsymbol{y}) - \nabla_{\boldsymbol{y}}\cdot\left(\Lambda\boldsymbol{f}(\boldsymbol{y})\right).$$

In (Raphan and Simoncelli, 2007a, 2010), we have developed a practical implementation of this NEBLS estimator, based on a local exponential approximation of the gradient of the log of the measurement density P_Y.

Additive Laplacian noise (scalar case). When the additive noise is drawn from a Laplacian distribution, we have

$$P_W(x) = \frac{1}{2\alpha}e^{-|x/\alpha|},$$

with Fourier transform

$$\widehat{P_W}(\omega) = \frac{1}{1+(\alpha\omega)^2},$$

which gives

$$\widehat{m}(\omega) = 2i\alpha^2\omega\widehat{P_W}(\omega).$$

The resulting BLS estimator is then

$$\mathbf{E}_{X|Y}\left(x|Y=y\right) = y + \frac{2\alpha^2 P'_W(y)\star P_Y(y)}{P_Y(y)}, \tag{33}$$

where $\star$ denotes convolution and

$$P'_W(y) = -(\frac{1}{\alpha})\mathrm{sgn}(y)\,P_W(y),$$

with

$$\mathrm{sgn}(y) = \begin{cases} -1, & y<0 \\ 0, & y=0 \\ 1, & y>0. \end{cases}$$

This solution uses a convolutional operator, as compared to the differentiation found in the Gaussian case. There are a variety of noise densities (for example, the family of generalized Gaussian distributions) for which the operator will be a convolution with a kernel that depends on the form of the noise. In these cases, the kernel may be used directly to approximate the convolutional operator from observed samples $\{y_k\}$:

$$K \star P_Y(y) \approx \frac{1}{N} \sum_{k=1}^{N} K(y - y_k).$$

Note that this has the form of a kernel density estimator. While such density estimators are generally biased (Scott, 1992), in our situation this approximation is unbiased and converges to the desired convolution $K \star P_Y$ as the number of samples (N) increases, since

$$\mathbf{E}_Y \left(\frac{1}{N} \sum_{k=1}^{N} K(y - y_k) \right) = \int K(y - \tilde{y}) P_Y(\tilde{y})\, d\tilde{y}.$$

Of course, the denominator of Eq. (33) still needs to be approximated using some choice of density estimator (see (Scott, 1992) for a review and further references).

Finally, substituting into Eq. (31) yields the dual operator:

$$\mathbf{L}^*\{f\}(y) = yf(y) - 2\alpha^2 (P_W' \star f)(y),$$

where the $\star$ indicates convolution.

Poisson process with random rate. Assume the hidden value, X, is positive and continuous, while the observation, Y, is discrete and has Poisson distribution with rate X:

$$P_{Y|X}(n|\, x) = \frac{e^{-x} x^n}{n!}.$$

It is easy to verify that

$$(n+1) P_{Y|X}(n+1|\, x) = x P_{Y|X}(n|\, x),$$

and from this, that

$$\mathbf{L}\{P_Y\}(n) = (n+1) P_Y(n+1).$$

As a result, the NEBLS estimator is

$$\mathbf{E}_{X|Y}\left(X|Y=n\right) = \frac{(n+1)P_Y(n+1)}{P_Y(n)},$$

which matches the original result derived by Robbins (1956).

Finally, the dual operator, $\mathbf{L}^*$, is readily derived from $\mathbf{L}$ by writing Eqs. (18a) and (18b), and replacing the integral by a sum:

$$\begin{aligned} \sum_{n=0}^{\infty} f(n)\, \mathbf{L}\{P_Y\}(n) &= \sum_{n=0}^{\infty} f(n)\, (n+1) P_Y(n+1) \\ &= \sum_{k=1}^{\infty} f(k-1)\, k P_Y(k) \\ &= \sum_{k=0}^{\infty} \mathbf{L}^*\{f\}(k)\, P_Y(k), \end{aligned}$$

from which we see that $\mathbf{L}^*\{f\}(n) = nf(n-1)$.

6.2 Non-invertible Observations

A NEBLS estimator always exists when the observation process, $\mathbf{A}$, is invertible, as can be seen from Eq. (13). In some cases (including some of those derived in this paper), the estimator exists even when $\mathbf{A}^{-1}$ is not defined. On the other hand, it is clear that some estimation problems do not allow a NEBLS form. Consider the extreme situation in which the observation contains no information about the quantity to be estimated: the optimal estimator is simply the mean of the prior density, which must be known in advance (i.e., it cannot be estimated from the data). In this section, we examine the conditions under which a NEBLS estimator may be defined for a non-invertible observation process.

We first decompose the prior into a sum of three orthogonal components, $P_X(\boldsymbol{x}) = P_1(\boldsymbol{x}) + P_2(\boldsymbol{x}) + P_3(\boldsymbol{x})$, with

$$\begin{aligned} P_1 &\in \mathcal{N}(\mathbf{A})^\perp \\ P_2 &\in \mathcal{N}(\mathbf{A}) \cap \mathcal{N}(\mathbf{A} \circ \mathbf{X})^\perp \\ P_3 &\in \mathcal{N}(\mathbf{A}) \cap \mathcal{N}(\mathbf{A} \circ \mathbf{X}), \end{aligned}$$

where $\mathcal{N}(\cdot)$ denotes the nullspace of an operator, and $()^\perp$ the orthogonal complement of a subspace. The measurement density may now be expressed as

$$P_Y = \mathbf{A} \circ P_X = \mathbf{A} \circ P_1,$$

since the second and third density components lie in the nullspace of $\mathbf{A}$. And since the first component of the prior is orthogonal to the nullspace, it may be recovered from the measurement density: $P_1 = \mathbf{A}^{\#}\{P_Y\}$, where $()^{\#}$ indicates the pseudo-inverse. Note that P_1 is guaranteed to integrate to one as long as P_Y does, since

$$\begin{aligned} \int P_Y(\boldsymbol{y})\, d\boldsymbol{y} &= \int \mathbf{A}\{P_1\}(\boldsymbol{y})\, d\boldsymbol{y} \\ &= \int\int P_{Y|X}(\boldsymbol{y}|\boldsymbol{x}) P_1(\boldsymbol{x})\, d\boldsymbol{x}\, d\boldsymbol{y} \\ &= \int \left[\int P_{Y|X}(\boldsymbol{y}|\boldsymbol{x})\, d\boldsymbol{y} \right] P_1(\boldsymbol{x})\, d\boldsymbol{x} \\ &= \int P_1(\boldsymbol{x})\, d\boldsymbol{x}. \end{aligned}$$

Substituting the decomposed prior into the numerator of the BLS estimator, as given by Eq. (12), produces

$$\begin{aligned} (\mathbf{A} \circ \mathbf{X})\{P_X\} &= (\mathbf{A} \circ \mathbf{X})\{P_1\} + (\mathbf{A} \circ \mathbf{X})\{P_2\} \\ &= \left(\mathbf{A} \circ \mathbf{X} \circ \mathbf{A}^{\#}\right)\{P_Y\} + (\mathbf{A} \circ \mathbf{X})\{P_2\} \\ &= \mathbf{L}\{P_Y\} + (\mathbf{A} \circ \mathbf{X})\{P_2\}, \end{aligned} \tag{34}$$

where we've discarded the term containing P_3 (since it lies in the nullspace of operator $\mathbf{A} \circ \mathbf{X}$, and replaced the term containing P_1 with its NEBLS equivalent. The second term depends on P_2, the component of the prior that cannot be recovered from the observation density but is nevertheless required to construct the BLS estimator. Thus, we can express the optimal estimator in NEBLS form if and only if the subspace containing this second component is zero (as is true of all solutions derived in this article). If not, then obtaining an optimal estimator requires *a priori* knowledge of that term.

Consider a simple example. Suppose we randomly select a coin with probability of heads $X \in [0,1]$, where X has prior density P_X. We then perform a binomial experiment, flipping the chosen coin n times and observing the number of heads, so that

$$P_{Y|X}\{k|x\} = \binom{n}{k} x^k (1-x)^{n-k},$$

where $\binom{n}{k} = \frac{n!}{(n-k)!k!}$. From this, we see that for each number of observed heads, k, the measurement probability consists of an inner product of P_X with a particular polynomial of degree n:

$$\begin{aligned} P_Y(k) &= \mathbf{A}\{P_X\}(k) \\ &= \binom{n}{k} \int_0^1 P_X(x) x^k (1-x)^{n-k}\, dx. \end{aligned} \tag{35}$$

Since the measurement process maps an arbitrary continuous prior density on the unit interval to a discrete measurement distribution, a simple dimensionality argument tells us that this process cannot be invertible. In particular, the measurement distribution contains the inner product of the prior density with $n+1$ linearly independent polynomials of degree n, so we can recover the inner product of our prior with any polynomial of degree n, but not with polynomials of higher degree. Define $\{q_k(x)\}_{k=0}^{\infty}$ as the set of orthogonal polynomials of ascending degree (obtained by starting with monomials and using Gram-Schmidt orthogonalization over the unit interval). Then $\mathcal{N}(\mathbf{A})^\perp$ is the space of polynomials of degree up to n, spanned by $\{q_k(x)\}_{k=0}^{n}$, and $\mathcal{N}(\mathbf{A})$ is spanned by $\{q_k(x)\}_{k=n+1}^{\infty}$. From the observation distribution, we may reconstruct P_1, the projection of the prior density onto $\mathcal{N}(\mathbf{A})^\perp$, but no component of the prior in $\mathcal{N}(\mathbf{A})$.

It might seem natural at this point to simply constrain the prior to be an nth order polynomial in X, which would allow recovery of the entire prior from the observation distribution. Although this constraint would certainly allow construction of a NEBLS estimator, it is far more restrictive than necessary. To construct the BLS estimator, we must be able to calculate its numerator, which is the inner product of the prior with a polynomial of degree $n+1$:

$$\begin{aligned} N(k) &= (\mathbf{A} \circ \mathbf{X})\{P_X\}(k) \\ &= \binom{n}{k} \int_0^1 P_X(x) x^{k+1} (1-x)^{n-k}\, dx \\ &= \binom{n}{k} \mathbf{E}_X \left(x^{k+1} (1-x)^{n-k} \right). \end{aligned}$$

This does not depend on the prior component P_3 which lies in the space spanned by $\{q_k(x)\}_{k=n+2}^{\infty}$, and therefore, there is no need to place any restrictions on this component of the prior (for example, by assuming it is equal to zero). The prior component P_2, on the other hand, which lies in the space spanned by $q_{n+1}(x)$ (i.e., $P_2(x) = c q_{n+1}(x)$, for some constant c), cannot be recovered from the observation density and *is* required to derive the BLS estimator. Thus, the BLS estimator may be written as a sum of a prior-free term, and a second term,

$$(\mathbf{A} \circ \mathbf{X})\{P_2\}(k) = c \binom{n}{k} \int_0^1 q_{n+1}(x) x^{k+1} (1-x)^{n-k}\, dx.$$

The value of c must be assumed *a priori*.

It is worth pointing out that this behavior is tied to the parameterization used. If, for example, we choose $X \in [0, \infty)$ with density P_X and then perform the Bernoulli experiment with probability of heads $\frac{X}{X+1}$ then

$$\begin{aligned} P_Y(k) &= \binom{n}{k} \int P_X(x) \left(\frac{x}{x+1}\right)^k \left(\frac{1}{x+1}\right)^{n-k} dx \\ &= \binom{n}{k} \int P_X(x) x^k \left(\frac{1}{x+1}\right)^n dx. \end{aligned}$$

In order to obtain the BLS estimator of X we need to know

$$N(k) = \binom{n}{k} \int P_X(x) x^{k+1} \left(\frac{1}{x+1}\right)^n dx.$$

Now it is easy to see that for $k < n$

$$N(k) = \frac{\binom{n}{k}}{\binom{n}{k+1}} P_Y(k+1).$$

Knowing $P_Y(k)$ gives the inner product of P_X, using weighting function $(\frac{1}{x+1})^n$, with all polynomials up to degree n. However, in order to know $N(n)$, we need to know the inner product of P_X with x^{n+1}. Therefore, in general we cannot solve for $N(n)$. Again, we can get around this by making assumptions about the missing (but necessary) portion of the prior.

Now consider the problem of developing a gSURE formula when the observation process is non-invertible. In this case, the constraints on the operator involve an interaction between the observation process and the family of estimators over which optimization occurs. From Eq. (18c), we see that the gSURE expression for the MSE relies on finding an operator $\mathbf{M}^*$ satisfying

$$\mathbf{E}_Y \left(\mathbf{M}^*\{\boldsymbol{f_\theta}\}(Y)\right) = \mathbf{E}_Y \left(\mathbf{E}_X \left(X|Y\right) \boldsymbol{f_\theta}(Y)\right), \qquad \forall \boldsymbol{f_\theta} \in \mathcal{F}, \tag{36}$$

which must hold for any observation density that could have arisen through the measurement process (i.e., $P_Y = \mathbf{A} \circ P_X$, for some density P_X). Decomposing the prior into orthogonal components as in Eq. (34), allows us to write

$$\mathbf{E}_{X|Y}\left(X|Y = \boldsymbol{y}\right) = \frac{\mathbf{L}\{P_Y\}(\boldsymbol{y})}{P_Y(\boldsymbol{y})} + \frac{(\mathbf{A} \circ \mathbf{X})\{P_2\}(\boldsymbol{y})}{P_Y(\boldsymbol{y})}.$$

Substituting this back into Eq. (36) and integrating by parts, we see that

$$\begin{aligned} \int \mathbf{M}^*\{\boldsymbol{f_\theta}\}(\boldsymbol{y})\, P_Y(\boldsymbol{y})\, d\boldsymbol{y} &= \mathbf{E}_Y\left(\frac{\mathbf{L}\{P_Y\}(Y)}{P_Y(Y)} \boldsymbol{f_\theta}(Y)\right) + \mathbf{E}_Y\left(\frac{(\mathbf{A} \circ \mathbf{X})\{P_2\}(Y)}{P_Y(Y)} \boldsymbol{f_\theta}(Y)\right) \\ &= \int \mathbf{L}^*\{\boldsymbol{f_\theta}\}(\boldsymbol{y})\, P_Y(\boldsymbol{y})\, d\boldsymbol{y} + \int (\mathbf{A} \circ \mathbf{X})\{P_2\}(\boldsymbol{y})\, \boldsymbol{f_\theta}(\boldsymbol{y})\, d\boldsymbol{y}, \qquad \forall \boldsymbol{f_\theta} \in \mathcal{F}, \end{aligned}$$

or, equivalently,

$$\int (\mathbf{M}^* - \mathbf{L}^*)\{\boldsymbol{f_\theta}\}(\boldsymbol{y})\, P_Y(\boldsymbol{y})\, d\boldsymbol{y} = \int (\mathbf{A} \circ \mathbf{X})\{P_2\}(\boldsymbol{y})\, \boldsymbol{f_\theta}(\boldsymbol{y})\, d\boldsymbol{y}, \qquad \forall \boldsymbol{f_\theta} \in \mathcal{F}, \tag{37}$$

which must hold for $P_Y = \mathbf{A}\{P_X\}$, for any prior $P_X = P_1 + P_2 + P_3$. Note that P_Y does not depend on the prior component P_2 (since it lies in the nullspace of $\mathbf{A}$). Thus, if we vary this component of the prior, the left side of Eq. (37) will stay the same while the right side will change, which implies that

$$\int (\mathbf{A} \circ \mathbf{X})\{P_2\}(\boldsymbol{y})\, \boldsymbol{f_\theta}(\boldsymbol{y})\, d\boldsymbol{y} = 0, \qquad \forall \boldsymbol{f_\theta} \in \mathcal{F}, \quad \forall P_2 \in \mathcal{N}(\mathbf{A}) \cap \mathcal{N}(\mathbf{A} \circ \mathbf{X})^{\perp}.$$

or equivalently

$$\mathcal{F} \subseteq (\mathbf{A} \circ \mathbf{X})\{P_2\}^{\perp}, \qquad \forall P_2 \in \mathcal{N}(\mathbf{A}) \cap \mathcal{N}(\mathbf{A} \circ \mathbf{X})^{\perp}. \tag{38}$$

This is therefore a necessary condition for the operator $\mathbf{M}^*$ to exist. It is also a sufficient condition, since if Eq. (38) is satisfied, then the operator $\mathbf{M} = \mathbf{L}$ will satisfy Eq. (37). Thus, selecting a family of estimators that satisfies the constraint of Eq. (38) guarantees that the optimal solution may be found through gSURE2PLS even when a NEBLS solution does not exist. Note that for over-restricted families of estimators, the choice of operator, $\mathbf{M}$, may not be unique.

In our coin tossing example, since the subspace of functions which are in the nullspace of $\mathbf{A}$ but orthogonal to the nullspace of $\mathbf{A} \circ \mathbf{X}$ is spanned by $q_{n+1}(x)$, Eq. (38) requires that

$$\sum_{k=0}^{n} \binom{n}{k} f_{\boldsymbol{\theta}}(k) \int q_{n+1}(x) x^{k+1} (1-x)^{n-k}\, dx = 0, \qquad \forall f_{\boldsymbol{\theta}} \in \mathcal{F},$$

(note that integral over y is replaced by a sum over k). Since, $q_{n+1}(x)$ is orthogonal to polynomials of degree less that $n+1$, this is equivalent to requiring that

$$\sum_{k=0}^{n} \binom{n}{k} f_{\boldsymbol{\theta}}(k) \int q_{n+1}(x) x^{n+1}\, dx = 0, \qquad \forall f_{\boldsymbol{\theta}} \in \mathcal{F},$$

which in turn implies that

$$\sum_{k=0}^{n} \binom{n}{k} f_{\boldsymbol{\theta}}(k) = 0, \qquad \forall f_{\boldsymbol{\theta}} \in \mathcal{F}.$$

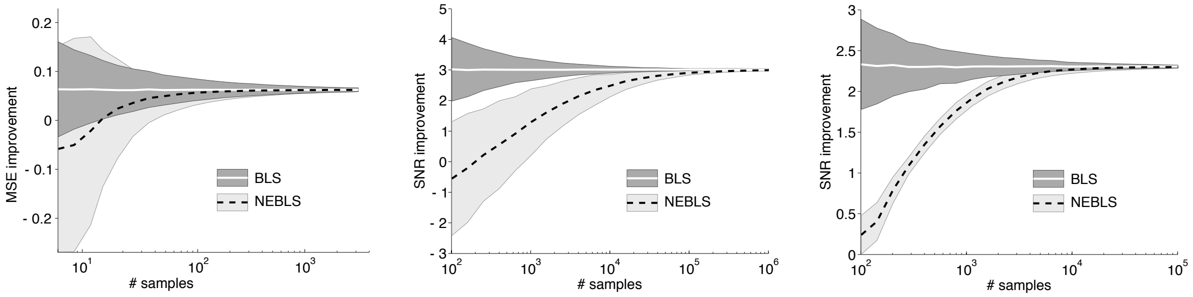
7 Empirical convergence properties

In this section, we implement several scalar NEBLS and gSURE2PLS estimators, and examine their convergence properties.

7.1 NEBLS examples

In practice, the NEBLS estimators rely on approximating the density of the observed data, $P_Y(Y)$, and the estimator quality depends critically on the choice of density estimator. If the density estimate converges to the true measurement density, then the associated NEBLS estimator should approach the BLS estimator as the number of data samples grows. In Fig. 3, we examine the behavior of three estimators based on Eq. (14). The first case corresponds to data drawn independently from a binary source, which are observed through a process in which bits are switched with probability $\frac{1}{4}$. The estimator does not know the binary distribution of the source (which was a “fair coin” for our simulation), but does know the bit-switching probability. For this estimator we use the observations to approximate P_Y using a simple histogram, and then use the matrix version of the linear operator in Eq. (13) to construct the estimator. We then apply the constructed estimator to the same observed data to estimate the uncorrupted value associated with each observation. We measure the behavior of the estimator, $\hat{x}(y)$, using the empirical MSE,

$$\frac{1}{N} \sum_{k=1}^{N} (\hat{x}(y_k) - x_k)^2, \tag{39}$$



(a) (b) (c)

Fig. 3: Empirical convergence of NEBLS estimator to conventional BLS solution, as a function of number of observed samples of Y. For each number of observations, each estimator is simulated many times. Black dashed lines show the improvement of the NEBLS estimator, averaged over simulations, relative to the ML estimator. White line shows the mean improvement using the conventional BLS solution, $\mathbf{E}_{X|Y}\left(X|Y=\boldsymbol{y}\right)$, assuming the prior density is known. Gray regions denote ± one standard deviation. (**a**) Binary noise (10,000 simulations for each number of samples); (**b**) Poisson noise (1,000 simulations); (**c**) additive Gaussian noise (1,000 simulations).

where $\{x_k\}$ are the underlying true values and $\{\hat{x}(y_k)\}$ are the corresponding estimates based on the observations. We characterize the behavior of this estimator as a function of the number of data points, N, by running many Monte Carlo simulations for each N, constructing the estimator using the N observations, applying the constructed estimator to these observations and recording the empirical MSE. Figure 3 indicates the mean improvement in empirical MSE (measured by the increase in empirical MSE compared with using the ML estimator, which, in this case, is the identity function) over the Monte Carlo simulations, the mean improvement using the conventional BLS estimation function, $\hat{x}(y) = \mathbf{E}_{X|Y}\left(X|Y=y\right)$, assuming the prior density is known, and the standard deviations of the improvements taken over our simulations. Note that the large variance in the BLS estimator for small numbers of data points arises from fluctuations of the empirical MSE.

Figure 3**b** shows the case of estimating a randomly varying rate parameter that governs an inhomogeneous Poisson process. The prior on the rate (unknown to the estimator) is exponential. The observed values Y are the (integer) values drawn from the Poisson process. In this case the histogram of observed data was used to obtain a naive approximation of $P_Y(n)$, the appropriate operator from Table 1 was used to convert this into an estimator, and this estimator was then applied to the observed data. It should be noted that improved performance for this estimator is expected if we were to use a more sophisticated approximation of the ratio of densities.

Figure 3**c** shows similar results for additive Gaussian noise, with the empirical MSE being replaced by the empirical Signal to Noise Ratio (SNR), which is defined as

$$SNR(dB) = 20\log_{10}\left(\frac{\sum_{k=1}^{N} x_k^2}{\sum_{k=1}^{N}(\hat{x}(y_k)-x_k)^2}\right). \tag{40}$$

The signal density is a generalized Gaussian with exponent 0.5, and the noisy SNR is 4.8 dB. In this case, we compute Eq. (14) using a more sophisticated approximation method, as described in (Raphan and Simoncelli, 2007a). We fit a local exponential model similar to that used in (Loader, 1996) to the data in bins, with binwidth adaptively selected so that the product of the number of points in the bin and the squared binwidth is constant. This binwidth selection procedure, analogous to adaptive binning procedures developed

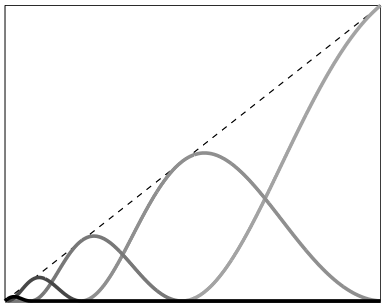


Fig. 4: Example "bump" kernel functions, as used for linear parameterization of "SUREbumps" estimators in Figs. 5(a) and 5(b). The sum of these functions is the identity (indicated by the dotted black line).

in the density estimation literature (Scott, 1992), provides a reasonable tradeoff between bias and variance, and converges to the correct answer for any well-behaved density (Raphan and Simoncelli, 2007a). Note that in this case, convergence is substantially slower than for the binary case, as might be expected given that we are dealing with a continuous density rather than a single scalar probability. But the variance of the estimates is quite low, even for relatively small amounts of data.

7.2 gSURE2PLS examples

Now consider the empirical behavior of the gSURE2PLS methodology in the additive Gaussian case, as developed in Sec. 2.2. The estimator is written as

$$f_{\boldsymbol{\theta}}(y) = y + g_{\boldsymbol{\theta}}(y),$$

and the parameter vector, $\boldsymbol{\theta}$, may be optimized over the observed data using the expression given by Eq. (8). As in the parameterization of Sec. 5, we write the function $g_{\boldsymbol{\theta}}(y)$ as a linear combination of nonlinear kernels

$$g_{\boldsymbol{\theta}}(y) = \sum_j \boldsymbol{\theta}_j h_j(y) = \boldsymbol{\theta}^T \boldsymbol{h}(y), \tag{41}$$

where $\boldsymbol{h}(y)$ is a vector with j^{th} component equal to kernel function $h_j(y)$. We define these kernels as

$$h_j(y) = y\ \cos^2\left(\frac{1}{\alpha}\text{sgn}(y)\log_2\left(|y|/\sigma + 1\right) - \frac{j\pi}{2}\right),$$

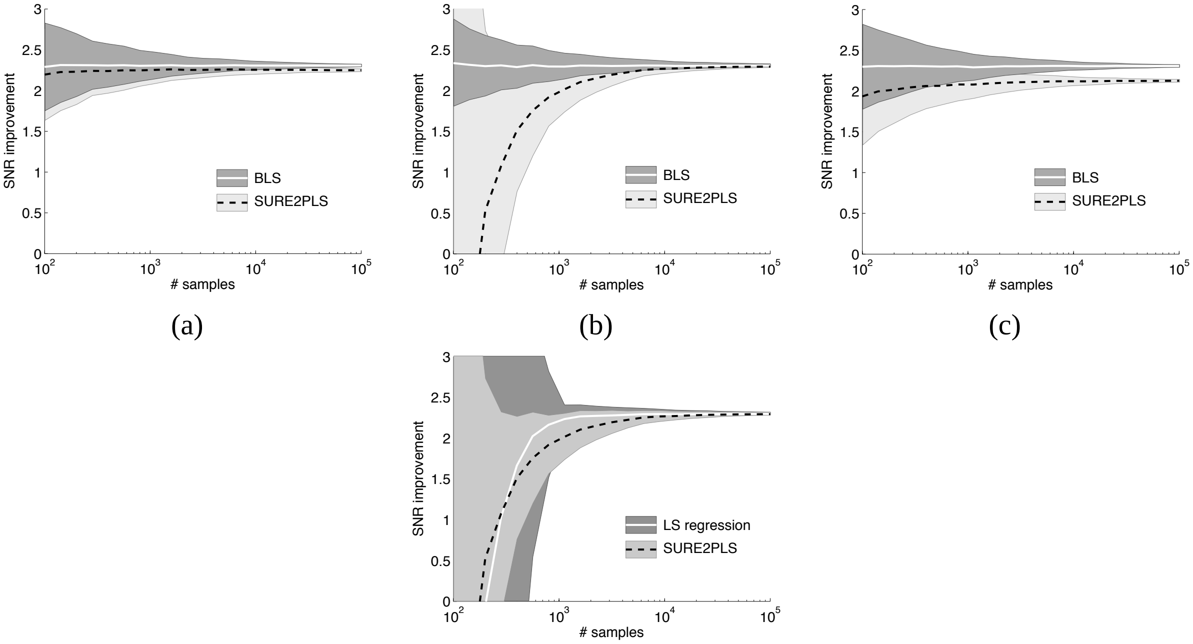
as illustrated in Fig. 4. Then, as in Sec. 5, substituting this into Eq. (7) yields a quadratic objective function with optimal solution

$$\hat{\boldsymbol{\theta}}_n = \mathbf{C}_n^{-1}\boldsymbol{m}_n, \tag{42}$$

where

$$\begin{aligned} \mathbf{C}_n &= \sum_{k=1}^{n} \boldsymbol{h}(y_k)\boldsymbol{h}(y_k)^T \\ \boldsymbol{m}_n &= -\sigma^2 \sum_{k=1}^{n} \boldsymbol{h}'(y_k). \end{aligned}$$

We apply this estimator to the same data that were used to obtain $\hat{\boldsymbol{\theta}}_n$, and measure the empirical SNR.



(a) (b) (c)

(d)

Fig. 5: Empirical convergence of gSURE2PLS methods, compared to optimal BLS solution, as a function number of data observations, for three different parameterized estimators. (**a**) 3-bump kernel estimator; (**b**) 15-bump kernel estimator; (**c**) Soft thresholding (SUREshrink). (**d**) Comparisons of gSURE2PLS and supervised regression, for the 15-bump kernel estimator. All simulations use a generalized Gaussian prior (exponent 0.5), and additive Gaussian noise. Noisy SNR is 4.8 dB.

For our simulations, we used a generalized Gaussian prior, with exponent 0.5. The noisy SNR was 4.8 dB. Figure 5 shows the empirical behavior of these “SUREbumps” estimators when using three bumps (Fig. 5**a**) and fifteen bumps (Fig. 5**b**), illustrating the bias-variance tradeoff inherent in the fixed parameterization. Three bumps behaves fairly well for small amounts of data, but the asymptotic behavior for large amounts of data is biased and thus falls short of ideal. Fifteen bumps asymptotes correctly but has very large variance for small amounts of data (i.e., it is overfitting). A more sophisticated method might use cross validation or some other resampling method to appropriately set the number of bumps to minimize both these effects. For comparison purposes, we have included the behavior of SUREshrink (Donoho and Johnstone, 1995), in which Eq. (7) is used to choose an optimal threshold, $\boldsymbol{\theta}$, for the function

$$f_{\boldsymbol{\theta}}(y) = \mathrm{sgn}(y)(|y| - \boldsymbol{\theta})^{+}.$$

As can be seen, SUREshrink shows significant asymptotic bias although the variance behavior is nearly ideal.

Figure 5(d) shows a comparison of the gSURE2PLS method to a supervised method, for the 15-bump kernel estimator (we find that the performance of the gSURE2PLS solution in the 3-bump and soft thresholding cases is nearly identical to that of the corresponding supervised regression solution). For each number of samples, N, the supervised method is trained on N pairs of state/measurement data, and then tested on a separate set of N measurements. Compared with the gSURE2PLS method, we see that the average performance of the supervised method converges slightly faster, but the variance is slightly worse.

8 Discussion

We have developed two general (and related) reformulations of the least squares estimation problem for the setting in which one knows the observation process, and has access to corrupted observations of many data samples. We do not assume knowledge of the prior density, nor do we assume access to samples from the prior.

The NEBLS form expresses the estimator in terms of a linear operator that depends only on the observation model. This unifies and generalizes those cases found in the literature (Robbins, 1956; Miyasawa, 1961; Maritz and Lwin, 1989). We discussed the conditions on measurement densities under which such NEBLS estimators may be obtained, provided a methodology for deriving them, and used it to derive solutions for a variety of specific observation situations. These estimators require one to estimate the observation density from observed data, and this may limit their usefulness, especially in high-dimensional cases. It is worth noting, however, that the NEBLS form can be employed successfully for the problem of denoising local blocks of signal or image data (Raphan and Simoncelli, 2010), where it provides an asymptotic correction of the recently developed “nonlocal means” estimator (Buades et al., 2005).

We also used the NEBLS form to express the MSE as an expectation over the measurement density. This provides a generalization of SURE and related methods for exponential cases (Stein, 1981; Berger, 1980; Hwang, 1982), as well as providing a new means of deriving them from their seemingly unrelated NEBLS counterparts. The generalized risk estimate may then be used to optimize a parametric estimator (dubbed the gSURE2PLS estimator), as has been done with SURE for the Gaussian noise case (Donoho and Johnstone, 1995; Pesquet and Leporini, 1997; Benazza-Benyahia and Pesquet, 2005; Luisier et al., 2006; Raphan and Simoncelli, 2007b; Blu and Luisier, 2007; Raphan and Simoncelli, 2008; Chaux et al., 2008). In cases where a NEBLS operator does not exist, we've shown how a parametric estimator family can be chosen so that gSURE2PLS is still valid. For any gSURE2PLS estimator that is parameterized as a sum of kernel functions, the gSURE optimization may be computed in closed form, and this leads naturally to an incremental algorithm in which the estimator is simultaneously refined by and applied to an incoming stream of measurements. Note that this incremental solution is general, and not limited to the case of additive Gaussian observations. We also showed that the two forms may be combined to yield a new form of parametric density estimation that is equivalent to score matching when the observation process is additive Gaussian noise.

Finally, we have implemented several NEBLS and SURE2PLS estimators, and examined their empirical convergence properties, showing that these estimators can perform as well as their full Bayesian or supervised regression counterparts. The implementation of NEBLS estimators must be handled on a case by case basis, and can be tricky since it requires the estimation of the measurement density from observed samples. The gSURE2PLS case is generally more straightforward, with success depending primarily on selection of a parametric family that can provide a good approximation to the optimal estimator, and for which optimization of the gSURE2PLS objective function is feasible. More generally, one could imagine adjusting the complexity of the estimator family depending on the amount of data available (for example, using cross-validation methods), or incorporating prior information about a particular problem to regularize the solution.

We believe it should be possible to generalize and extend these methods further. For example, we've assumed squared error loss throughout, but it is worth considering whether other loss functions might allow some form of nonparametric empirical Bayes solution. Note that Eq. (15) implies that our formulation is easily extended to any loss function that can be expressed as MSE in a nonlinearly transformed signal space. The advantage of the BLS solutions is that they effectively smooth the prior with the likelihood before integrating, whereas other estimators (such as MAP) will not generally have this property. A recent methodology, known as the Discrete Universal Denoiser (DUDe) (Weissman et al., 2005), provides a method for computing an optimal unsupervised denoiser with arbitrary loss functions. But the algorithm relies on recovery of the entire prior from the observed data, and is thus restricted to discrete priors. Our other main assumption has been that the observation model is fully known, and we have not studied the effect of errors in this model on the performance of the resulting NEBLS and gSURE2PLS estimators. We believe it may also be possible to learn the dual operator $\mathbf{L}^*$ satisfying Eq. (18c) from supervised data (e.g., for a scenario in which the measurement process is unknown but stable, and the statistics of the environment are drifting, thus precluding the use of supervised regression). Note that in this case, it is only necessary to learn the operator $\mathbf{L}^*$ in terms of its action on members of the parametric estimator family.

More generally, we think that the framework described here could be relevant for the design and construction of machines that need to optimize their estimation behavior in an environmentally adaptive way. On the biological side, Bayesian inference has been used to explain a variety of phenomena in human sensory and motor behavior, but very little has been said about how these estimators can be implemented, and even less about how these estimators can be learned without supervision or built-in priors. The estimators discussed here may offer a promising avenue for resolving these issues.

A Incremental forms for gSURE2PLS estimators

In Sec. 5, we showed a simple incremental form of gSURE2PLS for an estimator written as a linear combination of kernel functions. Here, we expand on this, providing a more efficient and more general form.

First, we note that it may be desirable to weight the incremental update rule in Eq. (26):

$$\begin{aligned}\mathbf{C}_n &= a_n\mathbf{C}_{n-1} + (1-a_n)\boldsymbol{h}_n\boldsymbol{h}_n^T \\ \boldsymbol{m}_n &= b_n\boldsymbol{m}_{n-1} + (1-b_n)\mathbf{L}^*\{\boldsymbol{h}\}(y_n),\end{aligned}$$

where the weights, a_n and b_n, are scalars in the range $(0,1)$, and $\boldsymbol{h}_n$ is an abbreviation for $\boldsymbol{h}(y_n)$. The weighting can provide numerical stability (so that the stored quantities do not continue to grow indefinitely), and allow the estimator to adapt to slowly time-varying statistics. A value of $\frac{n-1}{n}$ will equally weight all past data, while smaller weights will weight recent data more heavily. The choice of weighting allows a compromise between including more data (which reduces the variance of the estimation error) and adapting more rapidly (which reduces bias). The former depends on the complexity/dimensionality of the parameterization, and the latter depends on the rate at which the environmental statistics change.

Second, note that the incremental solution as provided in Sec. 5 requires the inversion of a matrix, which can be expensive, depending on the number of parameters. As is common in the derivation of the Kalman filter, we can use the Woodbury matrix identity (Hager, 1989) to rewrite the incremental form directly in terms of the inverse matrix:

$$\begin{aligned}\mathbf{C}_n^{-1} &= \left(a_n\mathbf{C}_{n-1} + (1-a_n)\boldsymbol{h}_n\boldsymbol{h}_n^T\right)^{-1} \\ &= a_n^{-1}\mathbf{C}_{n-1}^{-1} - a_n^{-1}\mathbf{C}_{n-1}^{-1}\boldsymbol{h}_n\left((1-a_n)^{-1} + \boldsymbol{h}_n^T a_n^{-1}\mathbf{C}_{n-1}^{-1}\boldsymbol{h}_n\right)^{-1}\boldsymbol{h}_n^T a_n^{-1}\mathbf{C}_{n-1}^{-1} \\ &= a_n^{-1}\left[\mathbf{C}_{n-1}^{-1} - \left(\frac{a_n}{1-a_n} + \boldsymbol{h}_n^T\boldsymbol{v}_n\right)^{-1}\boldsymbol{v}_n\boldsymbol{v}_n^T\right],\end{aligned}$$

where we have defined

$$\boldsymbol{v}_n \equiv \mathbf{C}_{n-1}^{-1}\boldsymbol{h}_n.$$

Note that since $\boldsymbol{h}_n^T\boldsymbol{v}_n$ is a scalar, computation of this expression does not require matrix inversion. Putting all of this together, and letting $\mathbf{S}_n$ denote $\mathbf{C}_n^{-1}$, the incremental algorithm is defined by the following set of equations:

$$\boldsymbol{v}_n = \mathbf{S}_{n-1}\boldsymbol{h}_n \tag{44a}$$

$$\mathbf{S}_n = a_n^{-1}\left[\mathbf{S}_{n-1} - \left(\frac{a_n}{1-a_n} + \boldsymbol{h}_n^T\boldsymbol{v}_n\right)^{-1}\boldsymbol{v}_n\boldsymbol{v}_n^T\right] \tag{44b}$$

$$\boldsymbol{m}_n = b_n\boldsymbol{m}_{n-1} + (1-b_n)\mathbf{L}^*\{\boldsymbol{h}\}(y_n) \tag{44c}$$

$$\hat{\boldsymbol{\theta}}_n = \mathbf{S}_n\boldsymbol{m}_n \tag{44d}$$

$$\hat{x}_n = \boldsymbol{h}_n^T\hat{\boldsymbol{\theta}}_n. \tag{44e}$$

The matrix $\mathbf{S}_n$ and the vector $\boldsymbol{m}_n$ constitute the stored state variables, and $\boldsymbol{v}_n$, $\hat{\boldsymbol{\theta}}_n$, and $\hat{x}_n$ are calculated based on this state and the observed data, y_n (or, more specifically, the observed data processed by the kernels, $\boldsymbol{h}(y_n)$ and $\mathbf{L}^*\{\boldsymbol{h}\}(y_n)$).

Finally, it is also possible to rewrite these equations so that the parameter vector, $\hat{\boldsymbol{\theta}}_n$, takes the role of the stored state variable, in place of $\boldsymbol{m}_n$. Specifically, we can substitute Eq. (44c) into Eq. (44d) to obtain:

$$\begin{aligned}
\hat{\boldsymbol{\theta}}_n &= b_n \mathbf{S}_n \boldsymbol{m}_{n-1} + (1-b_n)\mathbf{S}_n \mathbf{L}^*\{\boldsymbol{h}\}(y_n) \\
&= \frac{b_n}{a_n}\left[\mathbf{S}_{n-1} - \left(\frac{a_n}{1-a_n} + \boldsymbol{h}_n^T \boldsymbol{v}_n\right)^{-1} \boldsymbol{v}_n \boldsymbol{v}_n^T\right] \boldsymbol{m}_{n-1} + (1-b_n)\mathbf{S}_n \mathbf{L}^*\{\boldsymbol{h}\}(y_n) \\
&= \frac{b_n}{a_n}\left[\hat{\boldsymbol{\theta}}_{n-1} - \left(\frac{a_n}{1-a_n} + \boldsymbol{h}_n^T \boldsymbol{v}_n\right)^{-1} \boldsymbol{v}_n \boldsymbol{h}_n^T \hat{\boldsymbol{\theta}}_{n-1}\right] + (1-b_n)\mathbf{S}_n \mathbf{L}^*\{\boldsymbol{h}\}(y_n).
\end{aligned}$$

B Derivations of additional NEBLS estimators

In this section, we provide derivations of the remainder of the NEBLS estimators and their dual operators, as listed in Table 1 (derivations for general additive noise and Poisson noise are in Sec. 6.1).

B.1 Discrete exponential families

First, consider the discrete exponential family of the form

$$P_{Y|X}(n|x) = h(x)g(n)x^n,$$

where h is chosen to normalize the density (i.e., so that summing over n gives one). This case includes the Poisson case discussed in the previous section, amongst others. Noting that

$$P_{Y|X}(n+1|x) = h(x)g(n+1)x^{n+1},$$

we see by inspection that an operator satisfying the eigenfunction relationship of Eq. (27) is

$$\mathbf{L}\{P_Y\}(n) = \frac{g(n)}{g(n+1)} P_Y(n+1),$$

which tells us that

$$\mathbf{E}_{X|Y}\left(X|Y=n\right) = \frac{g(n)}{g(n+1)} \cdot \frac{P_Y(n+1)}{P_Y(n)},$$

as found in (Maritz and Lwin, 1989). The dual form is readily obtained from Eqs. (18a) and (18b) :

$$\mathbf{L}^*\{f\}(n) = \frac{g(n-1)}{g(n)} f(n-1).$$

Also, we note from Eq. (15) that the appropriate linear operator for $\mathbf{E}_{X|Y}\left(\frac{1}{X}|Y=n\right)$ will be

$$\mathbf{L}\{P_Y\}(n) = \frac{g(n)}{g(n-1)} P_Y(n-1).$$

This means that if $P_{Y|X}$ is instead parameterized as

$$P_{Y|X}(n|x) = h(x)g(n)x^{-n},$$

we will have

$$\mathbf{E}_{X|Y}\left(X|Y=n\right) = \frac{g(n)}{g(n-1)} \cdot \frac{P_Y(n-1)}{P_Y(n)}.$$

The dual form is then

$$\mathbf{L}^*\{f\}(n) = \frac{g(n+1)}{g(n)} f(n+1).$$

B.2 Continuous exponential families

Now consider the continuous exponential family of the form

$$P_{Y|X}(y|x) = h(x)g(y)e^{T(y)x},$$

where we assume that T is differentiable. By inspection, we obtain

$$\mathbf{L}\{P_{Y|X}(y|x)\} = \frac{g(y)}{T'(y)} \frac{d}{dy}\left\{\frac{P_{Y|X}(y|x)}{g(y)}\right\} = xP_{Y|X}(y|x)$$

which gives the NEBLS estimator

$$\begin{aligned}\mathbf{E}_{X|Y}\left(X|Y=y\right) &= \frac{g(y)\frac{d}{dy}\{\frac{P_Y(y)}{g(y)}\}}{T'(y)P_Y(y)} \\ &= \frac{1}{T'(y)}\frac{d}{dy}\ln(\frac{P_Y(y)}{g(y)}).\end{aligned}$$

The dual operator is

$$\mathbf{L}^*\{f\}(y) = \frac{-1}{g(y)} \cdot \frac{d}{dy}\left(\frac{f(y)g(y)}{T'(y)}\right).$$

As before, we may also deduce that if the likelihood is instead parameterized as

$$P_{Y|X}(y|x) = h(x)g(y)e^{T(y)/x},$$

we then have[8]

$$\mathbf{L}\{P_Y\}(y) = -g(y)\int_y^\infty \frac{T'(\tilde{y})}{g(\tilde{y})}P_Y(\tilde{y})d\tilde{y},$$

and so

$$\mathbf{E}_{X|Y}\left(\frac{1}{X}|Y=y\right) = \frac{-g(y)\int_y^\infty \frac{T'(\tilde{y})}{g(\tilde{y})}P_Y(\tilde{y})d\tilde{y}}{P_Y(y)}.$$

The dual operator is then

$$\mathbf{L}^*\{f\}(y) = -\frac{T'(y)}{g(y)}\int_0^y g(\tilde{y})\ f(\tilde{y})\ d\tilde{y}.$$

[8] We are assuming here that y takes on positive values. In some cases y can take on negative, or both negative and positive values, in which case the limits of integration would have to be changed for the operator $\mathbf{L}$ or the dual operator $\mathbf{L}^*$.

A particular case is that of a Laplacian scale mixture, for which

$$P_{Y|X}(y|x) = \frac{1}{x} e^{-\frac{y}{x}}; \qquad x, y > 0,$$

so that

$$\mathbf{L}\{P_Y\}(y) = P_Y\{Y > y\},$$

and

$$\mathbf{E}_{X|Y}\left(X|Y = y\right) \;\;=\;\; \frac{P_Y\{Y > y\}}{P_Y(y)}.$$

B.3 Power of Fixed Density

An interesting family of observation processes are those for which

$$\widehat{P_{Y|X}}(\omega|x) = [\widehat{P_W}(\omega)]^x, \tag{45}$$

for some density P_W. This occurs, for example, when X takes on integer values, and Y is a sum of X i.i.d. random variables with distribution P_W. Taking the derivative of Eq. (45) gives

$$\begin{aligned}
\widehat{P_{Y|X}}'(\omega|x) &= \widehat{P_W}'(\omega)\, x\, \widehat{P_W}(\omega)^{x-1} \\
&= \frac{\widehat{P_W}'(\omega)}{\widehat{P_W}(\omega)}\, x\, \widehat{P_W}(\omega)^x \\
&= \frac{d}{d\omega} \ln(\widehat{P_W}(\omega))\, x\, \widehat{P_{Y|X}}(\omega|x).
\end{aligned}$$

Rearranging this equality, and using the fact that differentiation in the Fourier domain is multiplication by an imaginary ramp in the signal domain gives

$$\begin{aligned}
x\, \widehat{P_{Y|X}}(\omega|x) &= \frac{1}{\frac{d}{d\omega} \ln(\widehat{P_W}(\omega))} \widehat{P_{Y|X}}'(\omega|x) \\
&= \frac{1}{i \frac{d}{d\omega} \ln(\widehat{P_W}(\omega))} \widehat{yP_{Y|X}}(\omega|x).
\end{aligned}$$

and comparing to the desired eigenfunction relationship of Eq. (27) allows us to define the operator

$$\begin{aligned}
\mathbf{L}\{P_Y\}(y) &= \mathcal{F}^{-1}\left\{\frac{1}{i \frac{d}{d\omega} \ln(\widehat{P_W}(\omega))} \widehat{yP_Y}(\omega)\right\}(y) \\
&= \Big(m \star (yP_Y)\Big)(y), \qquad (46)
\end{aligned}$$

where

$$m(y) = \mathcal{F}^{-1}\left\{\frac{1}{i \frac{d}{d\omega} \ln(\widehat{P_W}(\omega))}\right\}. \tag{47}$$

Thus, the linear operation first multiplies P_Y by y and then convolves with $m(y)$. The corresponding dual operator must first convolve $f(y)$ with the dual of $m(y)$, and then multiply by y, which we can express in the Fourier domain as:

$$\mathbf{L}^*\{f\}(y) = y \cdot \mathcal{F}^{-1}\{\widetilde{\widehat{m}(\boldsymbol{\omega})} \cdot \hat{f}(\boldsymbol{\omega})\}.$$

Four special cases are of particular interest. The first occurs when X is a positive variable and Y is a Poisson random variable with rate X. This corresponds to Eq. (45) with

$$\widehat{P_W}(\omega) = e^{(e^{-i\omega}-1)}.$$

Substituting into Eq. (47) gives

$$\widehat{m}(\omega) = e^{i\omega}.$$

Substituting this into Eq. (46), taking the inverse Fourier transform and substituting into Eq. (10), gives the estimator

$$\mathbf{E}_{X|Y}\left(X|Y=n\right) = \frac{(n+1)P_Y(n+1)}{P_Y(n)},$$

consistent with the result derived in Sec. 6.1.

The second example arises when X is a positive random variable and Y is a zero mean Gaussian with variance X, a case known as the Gaussian Scale Mixture (GSM) (Andrews and Mallows, 1974). In this case Eq. (45) holds for

$$\widehat{P_W}(\omega) = e^{-\frac{1}{2}\omega^2},$$

and the operator will be

$$\widehat{m}(\omega) = \frac{-1}{i\omega},$$

which gives

$$\mathbf{E}_{X|Y}\left(X|Y=y\right) = \frac{-(H(y)-\frac{1}{2})\star(yP_Y(y))}{P_Y(y)},$$

where H is the Heaviside step function. Since $yP_Y(y)$ is odd this is equal to

$$\begin{aligned}\mathbf{E}_{X|Y}\left(X|Y=y\right) &= \frac{-(H(y))\star(yP_Y(y))}{P_Y(y)}\\ &= \frac{-\int_{-\infty}^{y}\tilde{y}P_Y(\tilde{y})d\tilde{y}}{P_Y(y)}\\ &= \frac{-E_Y\{Y;Y<y\}}{P_Y(y)},\end{aligned}$$

where the numerator is now the mean of the density to the left of y and may be approximated in an unbiased way by the average of data less than y. Note that since $E_Y\{Y\}=0$, this may also be written as [9]

$$\mathbf{E}_{X|Y}\left(X|Y=y\right) = \frac{E_Y\{Y;Y>y\}}{P_Y(y)},$$

in agreement with our results in Section B.2.

A third special case occurs when $Y \sim \mathcal{N}(aX, X)$. That is, Y is aX corrupted by by zero mean additive Gaussian noise, with variance equal to X (we can trivially generalize to variance which is linear in X). In this case Eq. (45) will hold for

$$\widehat{P_W}(\omega) = e^{-\frac{1}{2}\omega^2 - ia\omega}.$$

[9] More generally, it may be the case that more than one operator $\mathbf{L}$ agree on the range of $\mathbf{A}$ (i.e., they agree on all possible P_Y that may arise from a particular process). In this case, the NEBLS estimator may not be unique.

In this case, the NEBLS operator is

$$\mathbf{L}\{P_Y\}(y) \quad = \quad \left(\mathcal{F}^{-1}\{\frac{1}{a - i\omega}\} \star (yP_Y)\right)(y),$$

where

$$\mathcal{F}^{-1}\{\frac{1}{a - i\omega}\}(y) \quad = \quad \text{sgn}(a)e^{ay}I_{\{ay<0\}},$$

so that

$$\mathbf{E}_{X|Y}\left(X|Y = y\right) = \frac{\left((\text{sgn}(a)e^{ay}I_{\{ay<0\}}) \star (yP_Y)\right)(y)}{P_Y(y)}.$$

A fourth special case is when X is a random positive value, W is an independent variable drawn from an α-stable distribution (Feller, 1970) with Fourier transform

$$\widehat{P_W}(\omega) = e^{-\frac{1}{\alpha}|\omega|^\alpha},$$

and

$$Y = X^{\frac{1}{\alpha}}W.$$

Generally, if P_W is an infinitely divisible distribution and X is an arbitrary positive real number, then the right side of Eq.(45) will be the Fourier transform of some density, which can be used as the observation process. In the particular case of the alpha-stable distribution, the NEBLS operator is

$$\mathbf{L}\{P_Y\}(y) \quad = \quad \left(\mathcal{F}^{-1}\{\frac{i}{\text{sgn}(\omega)|\omega|^{\alpha-1}}\} \star (yP_Y)\right)(y)$$

So that

$$\mathbf{E}_{X|Y}\left(X|Y = y\right) = \frac{\left(\mathcal{F}^{-1}\{\frac{i}{\text{sgn}(\omega)|\omega|^{\alpha-1}}\} \star (yP_Y)\right)(y)}{P_Y(y)}.$$

B.4 Multiplicative Lognormal Noise

Now consider the case of multiplicative lognormal noise:

$$Y = Xe^W,$$

where W is Gaussian noise of variance σ^2, and independent of X. In this case, taking logarithms gives

$$\ln(Y) = \ln(X) + W,$$

yielding an additive Gaussian noise model. From the NEBLS solution for additive Gaussian noise (Eq. (32)), we have

$$\mathbf{E}_{X|Z}\left(\ln(X)|Z = z\right) = \frac{(z + \sigma^2 D_z)\{P_Z\}(z)}{P_Z(z)},$$

where $Z = \ln(Y)$ and D_z represents the derivative operator with respect to z. However, we wish to find $\mathbf{E}_{X|Y}(X|Y)$ so we need to use the change of variables formula in Eq. (16). Since $X = e^{\ln(X)}$, we have

$$\mathbf{E}_{X|Z}(X|Z=z) = \frac{e^{(z+\sigma^2 D_z)}\{P_Z\}(z)}{P_Z(z)}.$$

By the Baker-Campbell-Hausdorff formula (Wilcox, 1967) we have that

$$\begin{aligned} e^{(z+\sigma^2 D_z)}\{P_Z\}(z) &= e^{z+\frac{1}{2}\sigma^2}(e^{\sigma^2 D_z}\{P_Z\}(z)) \\ &= e^{z+\frac{1}{2}\sigma^2} P_Z(z+\sigma^2), \end{aligned}$$

so that

$$\mathbf{E}_{X|Z}(X|Z=z) = \frac{e^{z+\frac{1}{2}\sigma^2} P_Z(z+\sigma^2)}{P_Z(z)}.$$

Next, using the fact that $\ln(Y) = Z$, we have by the change of variables formula

$$P_Z(\ln(y)) = y P_Y(y),$$

so that

$$\mathbf{E}_{X|Y}(X|Y=y) = \frac{e^{\frac{3}{2}\sigma^2} P_Y(e^{\sigma^2} y)}{P_Y(y)} y.$$

Note that it would have been difficult to garner this result by simple inspection of the likelihood,

$$P_{Y|X}(y|x) = \frac{1}{y\sqrt{2\pi\sigma^2}}\, e^{-\frac{1}{2\sigma^2}(\ln(y/x))^2}.$$

The dual form is

$$\mathbf{L}^*\{f\}(y) = e^{-\sigma^2/2} y \cdot f(e^{-\sigma^2} y).$$

B.5 Mixture of Uniform Noise

Consider the case when our observation is drawn from a uniform density, whose width is controlled by hidden variable X:

$$P_{Y|X}(y|x) = \begin{cases} \frac{1}{2x}, & |y| \le x \\ 0, & |y| > x \end{cases},$$

where $x \ge 0$. The density of Y is thus a mixture of uniform densities. We note that the (complement of the) cumulative distribution is

$$\begin{aligned} \int_{|y|}^{\infty} P_{Y|X}(\tilde{y}|x) d\tilde{y} &= \begin{cases} \frac{1}{2x}(x - |y|), & |y| \le x \\ 0, & |y| > x \end{cases} \\ &= (x - |y|) P_{Y|X}(y|x), \end{aligned}$$

and thus, by inspection, an operator that has $P_{Y|X}$ as an eigenfunction may be written

$$L\{P_Y\}(y) = \int_{|y|}^{\infty} P_Y(\tilde{y}) d\tilde{y} + |y| P_Y(y),$$

giving

$$\begin{aligned}\mathbf{E}_{X|Y}\left(X|Y=y\right) &= |y| + \frac{\int_{|y|}^{\infty} P_Y(\tilde{y})d\tilde{y}}{P_Y(y)} \\ &= |y| + \frac{Pr\{Y > |y|\}}{P_Y(y)} \\ &= |y| + \frac{1 - Pr\{Y \leq |y|\}}{P_Y(y)}.\end{aligned}$$

That is, the estimator adds to the observed value the complement of the cumulative distribution divided by the measurement density. The dual operator is

$$\mathbf{L}^*\{f\}(y) = \begin{cases} |y|f(y) + \int_{-y}^{y} f(\tilde{y})\, d\tilde{y}, & y \geq 0, \\ |y|f(y), & y < 0. \end{cases}$$

Acknowledgments

We are grateful to Sam Roweis, Maneesh Sahani, and Yair Weiss for helpful comments on this work, and to the anonymous reviewers for their suggestions in improving the manuscript. This research was funded by the Howard Hughes Medical Institute, and by New York University through a McCracken Fellowship to MR.

References

Andrews, D. and Mallows, C. (1974). Scale mixtures of normal distributions. *J. Royal Stat. Soc.*, 36:99–102.

Benazza-Benyahia, A. and Pesquet, J. C. (2005). Building robust wavelet estimators for multicomponent images using Stein's principle. *IEEE Trans. Image Proc.*, 14(11):1814–1830.

Berger, J. (1980). Improving on inadmissible estimators in continuous exponential families with applications to simultaneous estimation of gamma scale parameters. *The Annals of Statistics*, 8:545–571.

Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer, 2nd edition.

Blu, T. and Luisier, F. (2007). The SURE-LET approach to image denoising. *IEEE Trans. Image Proc.*, 16(11):2778–2786.

Buades, A., Coll, B., and Morel, J. M. (2005). A review of image denoising algorithms, with a new one. *Multiscale Modeling and Simulation*, 4(2):490–530.

Carlin, B. C. and Louis, T. A. (2009). *Bayesian Methods for Data Analysis*. CRC Press, Chapman & Hall, Boca Raton, Fl, 3rd edition.

Casella, G. (1985). An introduction to empirical Bayes data analysis. *Amer. Statist.*, 39:83–87.

Chaux, C., Duval, L., Benazza-Benyahia, A., and Pesquet, J. (2008). A nonlinear Stein based estimator for multichannel image denoising. *IEEE Trans. Image Proc.*, 56(8):3855–3870.

Comaniciu, D. and Meer, P. (2002). Mean shift: A robust approach toward feature space analysis. *IEEE Pat. Anal. Mach. Intell.*, 24:603–619.

Cressie, N. (1982). A useful empirical Bayes identity. *The Annals of Statistics*, 10(2):625–629.

Donoho, D. and Johnstone, I. (1995). Adapting to unknown smoothness via wavelet shrinkage. *J American Stat Assoc*, 90(432):1200–1224.

Eldar, Y. C. (2009). Generalized SURE for exponential families: Applications to regularization. *IEEE Trans. on Signal Processing*, 57(2):471–481.

Feller, W. (1970). *An introduction to probability theory and its applications*, volume 2. Wiley.

Hager, W. W. (1989). Updating the inverse of a matrix. *SIAM Review*, 31:221–239.

Hwang, J. T. (1982). Improving upon standard estimators in discrete exponential families with applications to poisson and negative binomial cases. *The Annals of Statistics*, 10:857–867.

Hyvärinen, A. (2005). Estimation of non-normalized statistical models using score matching. *Journal of Machine Learning Research*, 6:695–709.

Hyvärinen, A. (2008). Optimal approximation of signal priors. *Neural Computation*, 20:3087–3110.

Kalman, R. (1960). A new approach to linear filtering and prediction problems. *Transactions of the American Society of Mechanical Engineers –Journal of Basic Engineering*, 82D:35–45.

Loader, C. R. (1996). Local likelihood density estimation. *Annals of Statistics*, 24(4):1602–1618.

Luisier, F., Blu, T., and Unser, M. (2006). SURE-based wavelet thresholding integrating inter-scale dependencies. In *Proc. IEEE Int'l Conf. Image Processing*, pages 1457–1460, Atlanta GA.

Maritz, J. S. and Lwin, T. (1989). *Empirical Bayes Methods*. Chapman & Hall, 2nd edition.

Miyasawa, K. (1961). An empirical Bayes estimator of the mean of a normal population. *Bull. Inst. Internat. Statist.*, 38:181–188.

Morris, C. N. (1983). Parametric empirical Bayes inference: Theory and applications. *J. American Statistical Assoc.*, 78:47–65.

Pesquet, J. C. and Leporini, D. (1997). A new wavelet estimator for image denoising. In *6th International Conference on Image Processing and its Applications*, pages 249–253, Dublin, Ireland.

Raphan, M. (2007). *Optimal estimation: Prior free methods and physiological application*. PhD thesis, Courant Institute of Mathematical Sciences, New York University, New York, NY. Recipient, K. O. Friedrichs Prize for an outstanding disseration in mathematics.

Raphan, M. and Simoncelli, E. P. (2007a). Empirical Bayes least squares estimation without an explicit prior. Technical Report TR2007-900, Computer Science Technical Report, Courant Inst. of Mathematical Sciences, New York University.

Raphan, M. and Simoncelli, E. P. (2007b). Learning to be Bayesian without supervision. In Schölkopf, B., Platt, J., and Hofmann, T., editors, *Adv. Neural Information Processing Systems 19*, volume 19, pages 1145–1152, Cambridge, MA. MIT Press.

Raphan, M. and Simoncelli, E. P. (2008). Optimal denoising in redundant representations. *IEEE Trans Image Processing*, 17(8):1342–1352.

Raphan, M. and Simoncelli, E. P. (2009). Learning least squares estimators without assumed priors or supervision. Technical Report TR2009-923, Computer Science Technical Report, Courant Inst. of Mathematical Sciences, New York University.

Raphan, M. and Simoncelli, E. P. (2010). An empirical Bayesian interpretation and generalization of nl-means. Technical Report TR2010-934, Computer Science Technical Report, Courant Inst. of Mathematical Sciences, New York University.

Robbins, H. (1956). An empirical Bayes approach to statistics. *Proc. Third Berkley Symposium on Mathematcal Statistics*, 1:157–163.

Scott, D. W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization*. John Wiley.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *Annals of Statistics*, 9(6):1135–1151.

Wasserman, L. (2004). *All of Statistics: A Concise Course in Statistical Inference*. Springer.

Weissman, T., Ordentlich, E., Seroussi, G., Verdú, S., and Weinberger, M. (2005). Universal discrete denoising: Known channel. *IEEE Trans. Info. Theory*, 51(1):5–28.

Wilcox, R. M. (1967). Exponential operators and parameter differentiation in quantum physics. *Journal of Mathematical Physics*, 8:962–982.