

Learning least squares estimators without assumed priors or supervision

Martin Raphan and **Eero P. Simoncelli**

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

Email: {*raphan, eero*}@*cns.nyu.edu*

August 7, 2009

Abstract

The two standard methods of obtaining a least-squares optimal estimator are (1) Bayesian estimation, in which one assumes a prior distribution on the true values and combines this with a model of the measurement process to obtain an optimal estimator, and (2) supervised regression, in which one optimizes a parametric estimator over a training set containing pairs of corrupted measurements and their associated true values. But many real-world systems do not have access to either supervised training examples or a prior model. Here, we study the problem of obtaining an optimal estimator given a measurement process with known statistics, and a set of corrupted measurements of random values drawn from an unknown prior. We develop a general form of nonparametric empirical Bayesian estimator that is written as a direct function of the measurement density, with no explicit reference to the prior. We study the observation conditions under which such “prior-free” estimators may be obtained, and we derive specific forms for a variety of different corruption processes. Each of these prior-free estimators may also be used to express the mean squared estimation error as an expectation over the measurement density, thus generalizing Stein’s unbiased risk estimator (SURE) which provides such an expression for the additive Gaussian noise case. Minimizing this expression over measurement samples provides an “unsupervised regression” method of learning an optimal estimator from noisy measurements in the absence of clean training data. We show that combining a prior-free estimator with its corresponding unsupervised regression form produces a generalization of the “score matching” procedure for parametric density estimation, and we develop an incremental form of learning for estimators that are written as a linear combination of nonlinear kernel functions. Finally, we show through numerical simulations that the convergence of these estimators can be comparable to their supervised or Bayesian counterparts.

1 Introduction

The problem of estimating signals based on partial, corrupted measurements arises whenever a machine or 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, forecasting, and machine learning, just to name a few. The two most well-known formulations of this general problem are *Bayesian estimation* and *supervised regression*.¹

The Bayesian methodology formulates the optimal solution in terms of three fundamental ingredients: a statistical model of the quantity to be estimated (the *prior* probability distribution); a statistical model of the measurement process (the *likelihood* function); and a function that assigns a cost to errors (the *loss* function). An optimal estimator is one that minimizes the expected loss over the joint distribution of values and measurements. Obtaining this estimator requires explicit and complete knowledge of these three ingredients. The most commonly used loss function is the squared error, for which the optimal estimators are known as Bayesian Least Squares (BLS) or Minimum Mean Square Error (MMSE) estimators. We will restrict ourselves to this case for the remainder of this article. The likelihood may be learned by a calibration of the observation device: before sending a machine out in to the world we can train it in some known environment so that it may learn exactly how its sensors corrupt data. But how does one obtain a prior distribution? The classical view is that the prior must come from innate or assumed knowledge of the environment in which the machine will operate.

Alternatively, the machine might be able to learn about the probabilistic structure of its environment through analysis of multiple corrupted measurements (a process generally known as *empirical Bayesian* estimation). Direct reconstruction of a prior distribution based on corrupted data is difficult if not impossible to achieve practically. For example, consider the simple case of additive Gaussian noise. In this case, the density of the observed measurements is equal to the convolution of the prior density with the Gaussian noise density, and inversion of this process (a deconvolution problem) is known to be unstable and computationally infeasible without some restriction on the set of prior densities. A common solution is to assume a parametric form for the prior (e.g., Gaussian) and then learn the parameters from the noisy observations. Most empirical Bayes estimators achieve this by maximizing likelihood, or matching moments, both of which are inconsistent with the primary goal of minimizing squared estimation error.

Now consider the regression methodology, which assumes no prior knowledge of the environment or of the corruption process. Instead, it relies on a set of *supervised* training data consisting of pairs of corrupted sensory observations and the clean attribute values from which they arise. The optimal estimator is then selected from a family of potential estimators, as the member that minimizes the empirical MSE over the training pairs. The most well-known example is linear regression, in which the optimal solution is obtained directly by solving a linear inverse problem. Once chosen, the machine can apply this estimator to subsequent unsupervised observations of the unknown environment. Of course, this assumes that set of clean values and the conditional distribution of the corrupted measurements given their clean counterparts are representative of those that will be encountered in the unknown environment.

¹Classically, the phrase “Bayesian estimation” is used in the context of estimating a single unknown value from a set of measurements, based on an assumed prior distribution, whereas “regression” refers to the problem of estimating multiple unknowns, each with one or more corrupted observations. In this paper, we will be addressing the latter case, even when using Bayesian terminology. As such, the “prior distribution” of clean values is imposed by nature, and the optimal Bayesian estimator should be based on this distribution, rather than assuming an arbitrary regularizing distribution.

If the environment is truly unknown, in the sense that we lack a prior probability description, and we cannot obtain clean samples, how can we train our machine? In this article, we develop two related answers to this question. In the empirical Bayesian context, we develop a “prior-free” expression for the least squares estimator directly in terms of the density of noisy measurements, which generalizes several specialized examples from previous literature [1, 2, 3]. In addition to unifying these results, our framework allows us to provide a complete characterization of observaton models for which a prior-free estimator exists, and to obtain specific solutions for a variety of additional corruption processes.

In the regression context, we show that the prior-free estimation form can be used to derive an expression for the MSE that is written as an expectation over the measurement density, again without reference to the prior. Again, specialized examples of this appear in the literature, including Stein’s unbiased risk estimator (SURE, which is derived for the case of additive Gaussian noise [4]), and a few other specific examples [5, 6]. In addition to unifying and generalizing these previous examples, our framework ties them directly to the seemingly unrelated prior-free methodology. In practice, approximating the reformulated MSE with a sample average (as has been done with SURE [7, 8, 9, 10, 11, 12, 13]) allows one to select an optimal parametric estimator based entirely on a set of corrupted measurements, a procedure we refer to generally as “unsupervised regression”. For the special case of an estimator paramaterized using a linear basis (i.e., a kernel estimator), we develop an incremental algorithm that simultaneously learns and applies an optimal estimator to a stream of incoming data samples. We also show that the prior-free solution may be combined with the unsupervised regression 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 [14, 15]. Finally, we compare the empirical convergence of several example prior-free and unsupervised estimators with their Bayesian or supervised counterparts. Preliminary versions of the work in this article were presented in [16, 10, 17].

2 Introductory example: Additive Gaussian noise

We begin with a simple scalar example. 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:

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

where the denominator contains the distribution of the observed data, which we refer to as the *measurement density* (it also known as 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 Bayesian estimation without an explicit prior

A remarkable result, originally discovered by Miyasawa [2], tells us that when the measurements are obtained from true values by adding independent Gaussian noise of variance σ^2 , the optimal estimator may be expressed entirely in terms of the measurement density. To see this, first note that the 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),

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

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

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

We refer to this form as a “prior-free Bayesian estimator”, since it is expressed directly as a function of the measurement density, with no explicit reference to the prior density.² The derivation relies only on the assumptions of the squared loss function, and on the additive Gaussian measurement noise. We will assume squared loss throughout this article, but in Sec. 3, we describe a prior-free 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 [18], which uses iterative gradient ascent on the logarithm of a density to perform nonparametric clustering.

²Technically, this is a form of *nonparametric empirical Bayesian estimator* [1]. We have introduced the “prior-free” terminology to distinguish it from the more common type of empirical Bayesian estimator, in which a parametric prior is explicitly stated, and the parameters of this prior are determined by optimizing some objective function on the observed data (typically, the likelihood). The prior-free terminology is also meant to emphasize the fact that (as we will show) these estimators serve as the basis for a larger framework for estimation without explicit prior information.

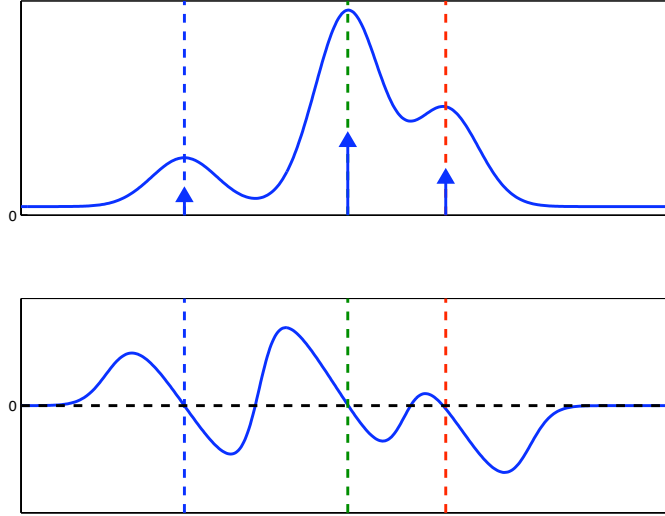


Fig. 1: Illustration of the prior-free estimator for a one-dimensional value with additive Gaussian noise. **Top:** Measurement density, $P_Y(y)$, arising from a signal with prior consisting of three point masses (indicated by upward arrows) corrupted by additive Gaussian noise. **Bottom:** Derivative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

2.2 Dual formulation: Regression without supervision

Next, as an alternative to the BLS estimator, consider the parametric regression formulation of the optimal estimation problem. Given a family of estimators, f_θ , parameterized by vector θ , we wish to select the one that minimizes the expected squared error:

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

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, $\{x_n, y_n\}$:

$$\hat{\theta} = \arg \min_{\theta} \frac{1}{N} \sum_n \left(f_\theta(y_n) - x_n \right)^2. \quad (5)$$

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

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

The middle term can be simplified by substituting the prior-free solution of Eq. (4), and then integrating by parts:

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

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

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

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

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 [4], who derived it³ and used it as a means of comparing the quality of different estimators. 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 by averaging over an unsupervised set of noisy measurements (i.e., with no reference to the clean signal values), and this approximation can then be minimized over the parameters, θ , to select an estimator:

$$\hat{\theta} = \arg \min_{\theta} \frac{1}{N} \sum_n \left(g_\theta^2(y_n) + 2\sigma^2 g'_\theta(y_n) \right), \tag{8}$$

where we've dropped the term σ^2 because it does not depend on the choice of parameter vector, θ . Since this solution does not rely on access to true signal values, $\{x_n\}$, (compare to the supervised regression form given by Eq. (5)), we refer to it as "unsupervised regression". This methodology for removal of additive Gaussian noise was first introduced by Donoho and Johnstone [7], who used it to estimate an optimal threshold

³Note that Stein derived the expression directly, without reference to Miyasawa's nonparametric Bayesian estimator. In addition, Stein formulated the problem in a 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*). Recent variants have been constructed using linear families of estimators [8, 9, 10, 11, 12, 13], which we discuss in Sec. 5.

Intuitively, this 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_n , 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.

We’ve developed the unsupervised regression solution by assuming a parametric estimator. Suppose instead that we start with a parametric form for the prior, $P_X^{(\phi)}(x)$. Combining this with the likelihood (using Eq. (3)) produces a parametric form for the measurement density, $P_Y^{(\phi)}(y)$. And given this, the prior-free expression of Eq. (4) can be used to define an optimal estimator: $g_\phi(y) = \sigma^2 \frac{d}{dy} \log P_Y^{(\phi)}(y)$. Finally, substituting this estimator into Eq. (8) and eliminating common factors of σ yields an objective function for the density parameters ϕ given samples $\{y_n\}$:

$$\hat{\phi} = \arg \min_{\phi} \frac{1}{N} \sum_n \left[\left(\frac{d}{dy} \log P_Y^{(\phi)}(y_n) \right)^2 + 2 \frac{d^2}{dy^2} \log P_Y^{(\phi)}(y_n) \right]. \quad (9)$$

The key point is that this objective function allows us to choose density parameters that are optimal for solving the least-squares estimation problem. By comparison, most empirical Bayes procedures select parameters for the prior density by optimizing some other criterion (e.g., maximizing likelihood of the data, or matching moments) [19], which are inconsistent with the estimation goal.⁴

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 [14], 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 [15], 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 [10]), 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 ML has been used to estimate parametric densities outside the context in which it is optimal (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.

⁴In fact, maximizing likelihood minimizes the Kullback-Leibler divergence between the true density and the parametric density [20].

3 General formulation: Prior-free BLS estimator

We now develop a generalization of the prior-free estimator of Eq. (4), and use it to derive a generalization of the prior-free MSE form of Eq. (7). Suppose we make a vector observation, \mathbf{y} , that is a corrupted version of an unknown vector \mathbf{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 as⁵

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

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

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

and rewrite the measurement density in terms of this operator:

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

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

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

where operator \mathbf{X} is defined as

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

Assume for the moment that \mathbf{A} is invertible, and that P_Y lies in the range of \mathbf{A} . Under these conditions, we can define \mathbf{A}^{-1} , the operator that inverts the observation process, recovering P_X from P_Y , and we can then write the numerator as a linear transformation on P_Y :

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

Note that in the discrete case, $P_Y(\mathbf{y})$ and $N(\mathbf{y})$ are each vectors, \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 prior-free form of the BLS estimator:

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

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

⁵The 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.

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 . This may seem like sleight of hand, given that our derivation assumed that the prior may be recovered exactly from the measurement density using \mathbf{A}^{-1} . First note that while the operator \mathbf{A}^{-1} may be ill-conditioned (e.g., a deconvolution in our introductory example), it is often the case that the composite operator \mathbf{L} is stable (e.g., a derivative in our introductory example). More surprisingly, even when \mathbf{A} is strictly non-invertible, it may be possible to find an operator \mathbf{L} that generates a prior-free estimator. The exact conditions under which the prior-free solution exists are derived in Sec. 6.2. Equation (14) provides a general form for the prior-free Bayesian estimator, and includes as special cases those formulations that have appeared in the literature on nonparametric empirical Bayes estimation (see Table 1 for specific citations).

The expression of Eq. (14) may be generalized to other prior-free expectations. For example, if we wished to calculate $E_{X|Y}\{X^n|Y = \mathbf{y}\}$, then Eq. (13) would be replaced by $(\mathbf{A} \circ \mathbf{X}^n \circ \mathbf{A}^{-1})\{P_Y\} = (\mathbf{A} \circ \mathbf{X} \circ \mathbf{A}^{-1})^n\{P_Y\} = \mathbf{L}^n\{P_Y\}$. Exploiting the linearity of the conditional expectation, we may extend this to *any* polynomial function (and thus to any function that can be approximated with a polynomial, $g(x) \approx \sum c_k x^k$):

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

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

$$r(Z) = Y,$$

with r an invertible, differentiable, transformation. Using known properties of change of variables for densities we have

$$\begin{aligned} \mathbf{E}_{X|Z}(X|Z = \mathbf{z}) &= \mathbf{E}_{X|Y}(X|Y = r(\mathbf{z})) \\ &= \frac{\mathbf{L}\{P_Y\}(r(\mathbf{z}))}{P_Y(r(\mathbf{z}))} \\ &= \frac{J_r(\mathbf{z})\mathbf{L}\{J_r^{-1}(r^{-1}(\mathbf{y}))P_Z(r^{-1}(\mathbf{y}))\}(r(\mathbf{z}))}{P_Z(\mathbf{z})}, \end{aligned} \tag{16}$$

where $J_r(\mathbf{z})$ is the Jacobian of the transformation, $r(\cdot)$.

4 General formulation: Unsupervised regression

As in the scalar Gaussian case, the prior-free 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 of estimators. This form is particularly useful in cases where it proves difficult to develop a stable nonparametric approximation of the ratio in Eq. (14).

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

$$\mathbf{E}_{X,Y} (|f_\theta(Y) - X|^2) = \mathbf{E}_{X,Y} (|f_\theta(Y)|^2 - 2f_\theta(Y) \cdot X + |X|^2). \quad (17)$$

Using the prior-free formulation of the previous section, the second term of the expectation may be written as

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

where \mathbf{L}^* is the dual operator of \mathbf{L} (in the discrete 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 θ), gives a prior-free expression for the optimal parameter vector:

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

This prior-free regression form includes as special cases those formulations that have appeared in previous literature⁶ (see 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 prior-free estimators. Conversely, it is relatively straightforward to show that the prior-free estimator of Eq. (14) can be derived from the prior-free regression expression of Eq. (19) (see [17] for a proof).

In practice, we can solve the optimal θ by minimizing the sample mean of this quantity:

$$\hat{\theta} = \arg \min_{\theta} \frac{1}{N} \sum_{n=1}^N \left\{ |f_\theta(y_n)|^2 - 2\mathbf{L}^* \{f_\theta\}(y_n) \right\}. \quad (20)$$

where $\{y_n\}$ 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, the unsupervised 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.

⁶As in the Gaussian (SURE) case, these previous results were described in a 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 3). 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.

It is worth noting that the estimator, $f_{\hat{\theta}}$, may be applied to the same data set that is used to optimize $\hat{\theta}$. 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 Sec. 5, we develop an algorithm for achieving this incrementally). In cases where one also has access to some supervised data $\{x_n, y_n | n \in \mathcal{S}\}$, in addition to unsupervised data $\{y_n | n \in \mathcal{U}\}$, the corresponding objective functions may be combined additively (since they both represent squared errors) to obtain a semi-supervised solution:

$$\begin{aligned}\hat{\theta} &= \arg \min_{\theta} \sum_{n \in \mathcal{S}} |f_{\theta}(y_n) - x_n|^2 + \sum_{n \in \mathcal{U}} \left\{ |f_{\theta}(y_n)|^2 - 2\mathbf{L}^*\{f_{\theta}\}(y_n) \right\} \\ &= \arg \min_{\theta} \sum_{n \in \mathcal{S} \cup \mathcal{U}} |f_{\theta}(y_n)|^2 - 2 \sum_{n \in \mathcal{S}} x_n f_{\theta}(y_n) - 2 \sum_{n \in \mathcal{U}} \mathbf{L}^*\{f_{\theta}\}(y_n),\end{aligned}$$

where we've again discarded a term that does not depend on θ . Again, the unsupervised regression 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. (18) can be generalized, using the result in Eq. (15), to give

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

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 by taking an expectation over Y :

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

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

Finally, the prior-free estimator for a parametric prior density may be substituted into the unsupervised objective function to obtain a generalized form of the score matching density estimator of Eq. (9). Specifically, the parametric density $P_Y^{(\phi)}$ may be fit to data $\{y_n\}$ by solving

$$\hat{\phi} = \arg \min_{\phi} \frac{1}{N} \sum_{n=1}^N \left[\left| \frac{\mathbf{L}\{P_Y^{(\phi)}\}}{P_Y^{(\phi)}}(y_n) \right|^2 - 2\mathbf{L}^* \left\{ \frac{\mathbf{L}\{P_Y^{(\phi)}\}}{P_Y^{(\phi)}} \right\}(y_n) \right]. \quad (21)$$

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} . But 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 Table 1), we obtain an estimator for parameter ϕ from data $\{n_k\}$:

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

5 Incremental unsupervised regression of kernel estimators

The unsupervised regression 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 [8, 9, 10, 11, 12, 13]. 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 an estimator that is formed as a weighted sum of fixed nonlinear kernel functions:

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

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

$$\hat{\theta} = \arg \min_{\theta} \left\{ \theta^T \left(\sum_{k=1}^n \mathbf{h}(y_k) \mathbf{h}(y_k)^T \right) \theta - 2\theta^T \sum_{k=1}^n \mathbf{L}^* \{ \mathbf{h} \} (y_k) \right\}, \quad (22)$$

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

$$\hat{\theta} = \mathbf{C}_n^{-1} \mathbf{m}_n, \quad (23)$$

where we define

$$\mathbf{C}_n \equiv \sum_{k=1}^n \mathbf{h}(y_k) \mathbf{h}(y_k)^T \quad (24a)$$

$$\mathbf{m}_n \equiv \sum_{k=1}^n \mathbf{L}^* \{ \mathbf{h} \} (y_k). \quad (24b)$$

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

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 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. (24) as:

$$\mathbf{C}_n = \mathbf{C}_{n-1} + \mathbf{h}(y_n) \mathbf{h}(y_n)^T \quad (25a)$$

$$\mathbf{m}_n = \mathbf{m}_{n-1} + \mathbf{L}^* \{ \mathbf{h} \} (y_n). \quad (25b)$$

corruption processes. But in many cases, it is difficult to obtain the operator \mathbf{L} directly from the prior-free estimation form of 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. (12) and (11) 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

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

That is, for each value of \mathbf{x} , the conditional density $P_{Y|X}(\mathbf{y}|\mathbf{x})$ is an *eigenfunction* (or eigenvector, for discrete variables) of operator \mathbf{L} , with associated eigenvalue \mathbf{x} . In many cases, we can use this relationship to obtain prior-free estimators by direct inspection of the observation density $P_{Y|X}$. Table 1 provides a listing of the examples which appear in the nonparametric empirical Bayes and Stein-related literatures, as well as a number of new ones we have derived from the more general framework presented here. Below, we provide a derivation for the case of general additive noise, and appendix B provides derivations of the others.

6.1 Additive noise: General case

Consider the case in which the variable of interest is corrupted by independent additive noise: $Y = X + W$, with the noise drawn from some distribution, $P_W(w)$. The conditional density may then be written

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

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

$$\mathbf{L}\{P_W(\mathbf{y} - \mathbf{x})\} = \mathbf{x}P_W(\mathbf{y} - \mathbf{x}). \quad (26)$$

Subtracting $\mathbf{y}P_W(\mathbf{y} - \mathbf{x})$ from both sides of Eq. (26) gives:

$$\mathbf{M}\{P_W(\mathbf{y} - \mathbf{x})\} = -(\mathbf{y} - \mathbf{x})P_W(\mathbf{y} - \mathbf{x}).$$

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

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

so that

$$\begin{aligned} \widehat{\mathbf{m}}(\omega) &= -i\frac{\nabla_\omega\widehat{P_W}(\omega)}{\widehat{P_W}(\omega)} \\ &= -i\nabla_\omega \ln\left(\widehat{P_W}(\omega)\right). \end{aligned} \quad (27)$$

	Obs. process	Obs. density: $P_{Y X}(y x)$	Numerator operator: $\mathbf{L}\{P_Y\}(y)$
	General discrete	\mathbf{A} (matrix)	$(\mathbf{A} \circ X \circ \mathbf{A}^{-1})P_Y(y)$
Additive	General	$P_W(y - x)$	yP_Y $-\mathcal{F}^{-1} \left\{ i \nabla_\omega \ln \left(\widehat{P_W}(\omega) \right) \widehat{P_Y}(\omega) \right\}$
	Gaussian [2]/[4]*	$\frac{\exp -\frac{1}{2}(y-x-\mu)^T \Lambda^{-1}(y-x-\mu)}{\sqrt{ 2\pi\Lambda }}$	$(y - \mu)P_Y(y) + \Lambda \nabla_y P_Y(y)$
	Poisson	$\sum \frac{\lambda^k e^{-\lambda}}{k!} \delta(y - x - ks)$	$yP_Y(y) - \lambda s P_Y(y - s)$
	Laplacian	$\frac{1}{2\alpha} e^{- y-x /\alpha}$	$yP_Y(y) + 2\alpha^2 \{P'_W \star P_Y\}(y)$
	Cauchy	$\frac{1}{\pi} \left(\frac{\alpha}{(\alpha(y-x))^2 + 1} \right)$	$yP_Y(y) - \left\{ \frac{1}{2\pi\alpha y} \star P_Y \right\}(y)$
	Uniform	$\begin{cases} \frac{1}{2a}, & y-x \leq a \\ 0, & y-x > a \end{cases}$	$yP_Y(y) + a \sum_k \text{sgn}(k) P_Y(y - ak)$ $-\frac{1}{2} \int P_Y(\tilde{y}) \text{sgn}(y - \tilde{y}) d\tilde{y}$
	Random # components	$P_W(y - x)$, where: $W \sim \sum_{k=0}^K W_k$, W_k i.i.d. (P_c) , $K \sim \text{Poiss}(\lambda)$	$yP_Y(y) - \lambda \{(yP_c) \star P_Y\}(y)$
	Gaussian scale mixture	$X + \sqrt{z}G$, $G \sim N(0, \Lambda)$	$yP_Y(y)$ $+\mathcal{F}^{-1} \left\{ \frac{\int_0^\infty z p_z(z) e^{-z \frac{1}{2} \omega^T \Lambda \omega} dz}{\int_0^\infty p_z(z) e^{-z \frac{1}{2} \omega^T \Lambda \omega} dz} \right\} \star \Lambda \nabla_y P_Y$
Cont. Exp.	Direct [3]/[5]*	$h(x)g(y)e^{T(y)x}$	$\frac{g(y)}{T'(y)} \frac{d}{dy} \left\{ \frac{P_Y(y)}{g(y)} \right\}$
	Inverse [5]*	$h(x)g(y)e^{T(y)/x}$	$g(y) \int_{-\infty}^y \frac{T'(\tilde{y})}{g(\tilde{y})} P_Y(\tilde{y}) d\tilde{y}$
	Laplacian scale mixture	$\frac{1}{x} e^{-\frac{y}{x}}; x, y > 0$	$P_Y\{Y > y\}$
Power of fixed	General	$\widehat{P_{Y X}}(\omega) = [\widehat{P_W}(\omega)]^X$	$\mathcal{F}^{-1} \left\{ \frac{1}{i \frac{d}{d\omega} \ln(\widehat{P_W}(\omega))} \right\} \star (yP_Y)$
	Gaussian scale mixture	$\frac{1}{\sqrt{2\pi x}} e^{-\frac{y^2}{2x}}$	$-E_Y\{Y; Y < y\}$
	Multiplicative α -stable	$Y = X^{\frac{1}{\alpha}} W, W \alpha$ -stable	$\mathcal{F}^{-1} \left\{ \frac{i}{\text{sgn}(\omega) \omega ^{\alpha-1}} \right\} \star (yP_Y(y))$
	Signal-dependent AWGN	$Y = aX + \sqrt{X}W, W \sim \mathcal{N}(0, 1)$	$\text{sgn}(a) e^{ax} I_{\{ax < 0\}} \star (yP_Y(y))$
Disc. Exp.	Direct [3]/[6]*	$h(x)g(n)x^n$	$\frac{g(n)}{g(n+1)} P_Y(n+1)$
	Inverse [6]*	$h(x)g(n)x^{-n}$	$\frac{g(n)}{g(n-1)} P_Y(n-1)$
	Poisson [1]/[6]*	$\frac{x^n e^{-x}}{n!}$	$(n+1)P_Y(n+1)$
	Uniform mixture	$\begin{cases} \frac{1}{2x}, & y \leq x \\ 0, & y > x \end{cases}$	$ y P_Y + Pr\{Y > y \}$
	Multiplicative lognormal	$Y = X e^W, W$ Gaussian	$e^{\frac{3}{2}\sigma^2} P_Y(e^{\sigma^2} y) y$

Table 1: Prior-free estimation formulas for a variety of observation processes, as listed in the left column. Bracketed numbers refer to 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 prior-free estimator, $\hat{x}_{\text{BLS}}(y) = \frac{\mathbf{L}\{P_Y\}(y)}{P_Y(y)}$. The symbol \star indicates convolution, and a hat (e.g., $\widehat{P_W}$) indicates a Fourier transform.

Substituting this into the definition of \mathbf{M} gives us the linear operator:

$$\mathbf{L}\{f\}(\mathbf{y}) = \mathbf{y} f(\mathbf{y}) - \mathcal{F}^{-1} \left\{ i \nabla_{\omega} \ln \left(\widehat{P}_W(\omega) \right) \widehat{f}(\omega) \right\}(\mathbf{y}), \quad (28)$$

where \mathcal{F}^{-1} denotes the inverse Fourier transform.

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 corruption process is not invertible (e.g., if the Fourier transform of P_W is bandlimited) the proof of Eq. (28) shows that this equation is still valid. We need simply define

$$\nabla_{\omega} \ln \left(\widehat{P}_W(\omega) \right) = 0,$$

whenever $\widehat{P}_W(\omega) = 0$. In this case, note that the observation process is not invertible, yet we are still able to define the prior-free operator \mathbf{L} . As examples, we consider the following specific cases of additive noise.

Additive Gaussian noise (vector case). The expression in Eq. (28) 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:

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

with covariance matrix Λ , and mean vector μ . In this case, the Fourier transform of the density is:

$$\widehat{P}_W(\omega) = e^{-i\omega \cdot \mu - \frac{1}{2}\omega^T \Lambda \omega},$$

which, upon substitution into Eq. (27) yields:

$$\widehat{m}(\omega) = [i\Lambda\omega - \mu] \widehat{P}_Y(\omega).$$

Computing the inverse Fourier transform and substituting into Eq. (28) yields

$$\begin{aligned} \mathbf{E}(X|Y=y) &= \mathbf{y} - \mu + \frac{\Lambda \nabla_{\mathbf{y}} P_Y(\mathbf{y})}{P_Y(\mathbf{y})} \\ &= \mathbf{y} - \mu + \Lambda \nabla_{\mathbf{y}} \ln(P_Y(\mathbf{y})). \end{aligned} \quad (29)$$

In [22], we have developed a practical implementation of this estimator, based on a local exponential approximation of the gradient of the log of the measurement density P_Y .

Additive Laplacian noise. 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

$$\mathbf{E}(X|Y = y) = y + \frac{2\alpha^2 P'_W(y) \star P_Y(y)}{P_Y(y)}, \quad (30)$$

where \star denotes convolution and

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

with

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

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 $\{Y_n\}$:

$$K \star P_Y(\mathbf{y}) \approx \frac{1}{N} \sum_{i=1}^N K(\mathbf{y} - Y_n).$$

Note that this has the form of a kernel density estimator. While such density estimators are generally biased [23], 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

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

Of course, the denominator of Eq. (30) still needs to be approximated using some choice of density estimator (see [23] for a review and further references).

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

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

and from this, that

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

As a result, the prior-free BLS estimator is

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

as previously derived in [1].

6.2 Non-invertible Observations

A prior-free 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 the inverse \mathbf{A}^{-1} is not defined. On the other hand, it is clear that some estimation problems do not allow a prior-free 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 prior-free estimator may be defined for a non-invertible observation process.

We first decompose the prior into a sum of three orthogonal components, $P_X(\mathbf{x}) = P_1(\mathbf{x}) + P_2(\mathbf{x}) + P_3(\mathbf{x})$, 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 $(\cdot)^\perp$ 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 $(\cdot)^\#$ indicates the pseudo-inverse. Note that P_1 is guaranteed to integrate to one as long as P_Y does, since

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

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

$$\begin{aligned} (\mathbf{A} \circ \mathbf{X}) \{P_X\} &= (\mathbf{A} \circ \mathbf{X}) \{P_1\} + (\mathbf{A} \circ \mathbf{X}) \{P_2\} \\ &= (\mathbf{A} \circ \mathbf{X} \circ \mathbf{A}^\#) \{P_Y\} + (\mathbf{A} \circ \mathbf{X}) \{P_2\} \\ &= \mathbf{L} \{P_Y\} + (\mathbf{A} \circ \mathbf{X}) \{P_2\}, \end{aligned} \tag{31}$$

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 by its prior-free 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 prior-free 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 (we can refer to such an estimator as "partially prior-free").

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

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

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 products of P_X with a particular polynomial of degree n :

$$\begin{aligned} P_Y(k) &= \mathbf{A}\{P_X\}(k) \\ &= \binom{n}{k} \int_0^1 P_X(x) x^k (1-x)^{n-k} dx. \end{aligned} \quad (32)$$

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 n th 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 prior-free 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 parametrization 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 formula for unsupervised regression when the observation process is noninvertible. 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. (18), we see that the prior-free expression for the MSE relies on finding an operator \mathbf{M}^* satisfying

$$\mathbf{E}_Y(\mathbf{M}^*\{f_\theta\}(Y)) = \mathbf{E}_Y(\mathbf{E}(X|Y) f_\theta(Y)), \quad \forall f_\theta \in \mathcal{F}, \quad (33)$$

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. (31), allows us to write

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

Substituting this back into Eq. (33) and integrating by parts, we see that

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

or, equivalently,

$$\int (\mathbf{M}^* - \mathbf{L}^*)\{f_\theta\}(y) P_Y(y) dy = \int (\mathbf{A} \circ \mathbf{X})\{P_2\}(y) f_\theta(y) dy, \quad \forall f_\theta \in \mathcal{F}, \quad (34)$$

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. (34) will stay the same while the right side will change, which implies that

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

or equivalently

$$\mathcal{F} \subseteq (\mathbf{A} \circ \mathbf{X})\{P_2\}^\perp, \quad \forall P_2 \in \mathcal{N}(\mathbf{A}) \cap \mathcal{N}(\mathbf{A} \circ \mathbf{X})^\perp. \quad (35)$$

This is therefore a necessary condition for the operator \mathbf{M}^* to exist. It is also a sufficient condition, since if Eq. (35) is satisfied, then the operator $\mathbf{M} = \mathbf{L}$ will satisfy Eq. (34). Thus, selecting a family of estimators that satisfies the constraint of Eq. (35) guarantees that the optimal solution may be found through unsupervised regression, even when a prior-free Bayesian solution does not exist. Note that for overrestricted families of estimators, the choice of dual 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. (35) requires that

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

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

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

which in turn implies that

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

7 Empirical examination of convergence properties

In this section, we demonstrate the feasibility of the prior-free and unsupervised regression methods by implementing several estimators, and examining their convergence properties.

7.1 Prior-Free BLS examples

In practice, the prior-free BLS estimators rely on approximating the density of the observed data, $P_Y(Y)$, and these estimators should approach the BLS estimator as the number of data samples grows, assuming the density approximation converges to the true measurement density. In Fig. 3, we examine the behavior of three non-parametric prior-free 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} , using the empirical MSE,

$$\frac{1}{N} \sum_{k=1}^N (\hat{X}_i - X_i)^2, \quad (36)$$

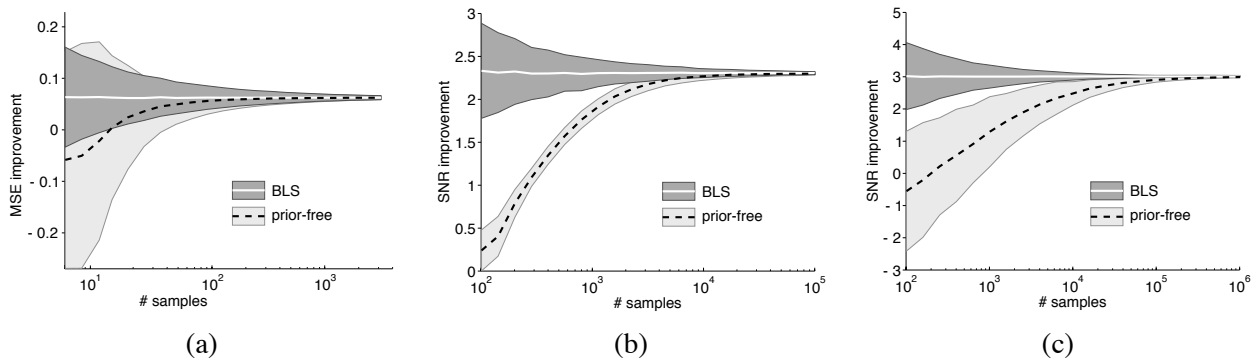


Fig. 3: Empirical convergence of prior-free estimator to optimal BLS solution, as a function number of observed samples of Y . For each number of observations, each estimator is simulated many times. Black dashed lines show the improvement of the prior-free estimator, averaged over simulations, relative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

where $\{X_i\}$ are the underlying values and $\{\hat{X}_i\}$ 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, $\mathbf{E}(X|Y = y)$ 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 3b 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

$$SNR(dB) = 20 \log_{10} \left(\frac{\sum_{k=1}^N X_k^2}{\sum_{k=1}^N (\hat{X}_k - X_k)^2} \right). \quad (37)$$

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 [22]. We fit a local exponential model similar to that used in [24] 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 in the density estimation literature [23], provides a reasonable tradeoff between bias and variance, and converges to the correct answer for any well-behaved density [22]. 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.

Figure 3c shows the case of estimating a randomly varying rate parameter that governs an inhomogeneous

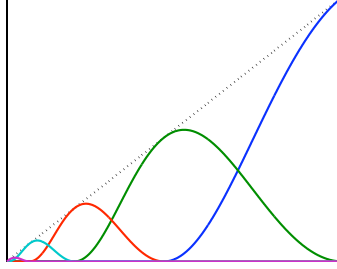


Fig. 4: Example “bump” kernel functions, as used for linear parameterization of estimators in Figs. 5(a) and 5(b). The sum of these functions is the identity (indicated by the dotted black line).

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.

7.2 Parametric unsupervised regression examples

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

$$f_\theta(y) = y + g_\theta(y),$$

and the parameter vector, θ , may be optimized over the observed data using the expression given by Eq. (8). Similar to the parameterization of Sec. 5, we write the function $g_\theta(y)$ as a linear combination of nonlinear “bump” kernels

$$g_\theta(y) = \sum_j \theta_j h_j(y) = \theta^T \mathbf{h}(y), \quad (38)$$

where $\mathbf{h}(y)$ is a vector with j^{th} component equal to function $h_j(y)$, defined as

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

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

$$\hat{\theta} = -\mathbf{C}_n^{-1} \mathbf{m}_n, \quad (39)$$

where

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

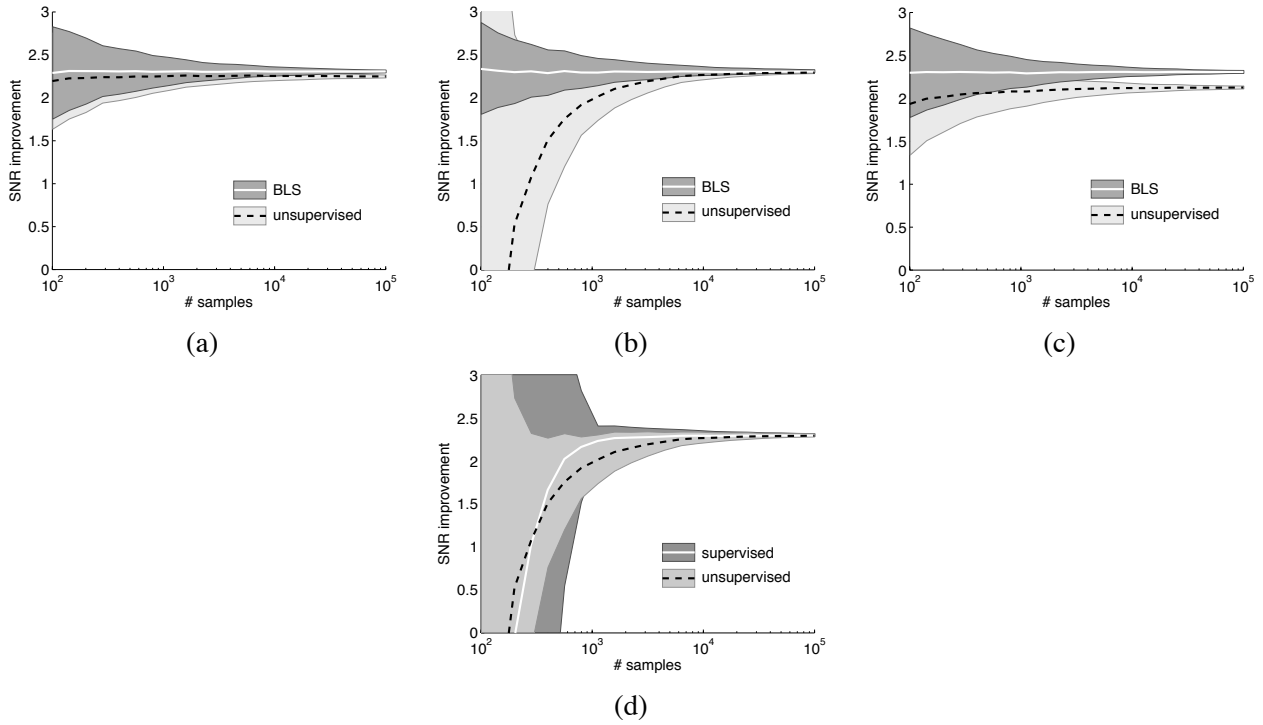


Fig. 5: Empirical convergence of unsupervised regression method compared to optimal BLS solution (a-c), as a function number of data observations, for three different parameterized estimators. (a) 3-bump kernel estimator; (b) 15-bump kernel estimator; (c) Soft thresholding. (d) Comparisons of unsupervised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

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

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 ‘‘SURE-bump’’ estimators when using three bumps (Fig. 5a) and fifteen bumps (Fig. 5b), 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 [7], in which Eq. (7) is used to choose an optimal threshold, θ , for the function

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

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

Figure 5(d) shows a comparison of the unsupervised regression method to a comparable supervised method, for the 15-bump kernel estimator (we only show this comparison for this case, because the supervised solution shows nearly identical performance to the unsupervised solution in the other two cases). 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. 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 (likelihood function), and has access to many corrupted observations. We do not assume knowledge of the prior density, nor do we assume access to samples from the prior. Our formulation thus acts as a bridge between a conventional Bayesian setting (in which one knows both the observation process and the prior), and the regression setting (in which one has samples from the prior, paired with corrupted observations of those samples). This formulation holds particular appeal for real-world systems (machine or biological) that must refine their estimates based on environmental observations.

The first form, which we have dubbed a ‘‘prior-free Bayesian estimator’’, expresses the estimator in terms of a linear operator that depends only on the observation model. This unifies and generalizes several special cases found in the statistics literature [2, 1, 3]. We also showed that this form may be extended to estimate arbitrary statistics of the unknown variable, or the expected value of any polynomial combination of the unknown and measurement variables, through an expectation over the measurement variable alone. We discussed the conditions on measurement densities under which prior-free estimators may be obtained, provided a methodology for deriving them, and used it to derive solutions for a variety of specific observation situations.

The second form, which we have dubbed ‘‘unsupervised regression’’, is derived by using the first form

to express the MSE as an expectation over the measurement density. This provides a generalization of SURE and related methods for exponential cases [4, 5, 6], as well as relating them with their prior-free counterparts. In cases where a prior-free operator does not exist, we've shown how a parametric estimator family can be chosen so that unsupervised regression is still valid. When the estimator is parameterized as a sum of kernel functions, the unsupervised regression solution may be computed in closed form, and leads naturally to an incremental algorithm in which the estimator is 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 prior-free estimators, and examined their empirical convergence properties, showing that in some cases, these estimators perform as well as their full Bayesian or supervised regression counterparts. The implementation of prior-free Bayesian estimators must be handled on a case by case basis, and can be tricky since it requires the estimation of the measurement density from the data. The unsupervised regression 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 unsupervised regression 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 prior-free solutions. 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 solution, known as the Discrete Universal Denoiser (DUDe) [25], 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 prior-free and unsupervised estimators. We believe it may also be possible to learn the dual operator \mathbf{L}^* satisfying Eq. (18) from data (rather than by assuming a known measurement process), by restricting the family of parametric estimators to be low-dimensional.

More generally, we think that the general 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 prior-free forms of these estimators discussed here may offer a promising avenue for resolving these issues.

A Incremental forms for unsupervised regression estimators

In Sec. 5, we showed a simple incremental form of unsupervised regression 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. (25):

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

where the weights, a_n and b_n , are scalars in the range $(0, 1)$, and \mathbf{h}_n is an abbreviation for $\mathbf{h}(y_n)$. 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 prior changes.

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 [26] to rewrite the incremental form directly in terms of the inverse matrix:

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

where we have defined

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

Note that since $\mathbf{h}_n^T \mathbf{v}_n$ 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

$$\mathbf{v}_n = \mathbf{S}_{n-1} \mathbf{h}_n \tag{41a}$$

$$\mathbf{S}_n = a_n^{-1} \left[\mathbf{S}_{n-1} - \left(\frac{a_n}{1 - a_n} + \mathbf{h}_n^T \mathbf{v}_n \right)^{-1} \mathbf{v}_n \mathbf{v}_n^T \right] \tag{41b}$$

$$\mathbf{m}_n = b_n \mathbf{m}_{n-1} + (1 - b_n) \mathbf{L}^* \{\mathbf{h}\}(y_n) \tag{41c}$$

$$\hat{\theta}_n = \mathbf{S}_n \mathbf{m}_n \tag{41d}$$

$$\hat{x}_n = \mathbf{h}_n^T \hat{\theta}_n. \tag{41e}$$

The matrix \mathbf{S}_n and the vector \mathbf{m}_n constitute the stored state variables, and \mathbf{v}_n , $\hat{\theta}_n$, 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, $\mathbf{h}(y_n)$ and $\mathbf{L}^* \{\mathbf{h}\}(y_n)$).

Finally, it is also possible to rewrite these equations so that the parameter vector, $\hat{\theta}_n$, takes the role of the stored state variable, in place of \mathbf{m}_n . Specifically, we can substitute Eq. (41c) into Eq. (41d) to obtain:

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

B Derivations of additional prior-free estimators

B.1 Mixture of Uniform Noise

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

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

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

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

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

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

giving

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

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

B.2 Multiplicative Lognormal Noise

Now consider the case of multiplicative lognormal noise:

$$Y = X e^W,$$

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 prior-free solution for additive Gaussian noise (Eq. (29)), we have

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

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

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

By the Baker-Campbell-Hausdorff formula [27] we have that

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

so that

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

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

$$P_Y(y) = \frac{P_Z(\ln(y))}{y},$$

so that

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

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

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

B.3 Power of Fixed Density

An interesting family of observation processes are those for which

$$\widehat{P}_{Y|X}(\omega) = [\widehat{P}_W(\omega)]^X, \tag{42}$$

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. (42) gives

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

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

$$\begin{aligned} x \widehat{P_{Y|X}}(\omega) &= \frac{1}{\frac{d}{d\omega} \ln(\widehat{P_W}(\omega))} \widehat{P_{Y|X}}'(\omega) \\ &= \frac{1}{i \frac{d}{d\omega} \ln(\widehat{P_W}(\omega))} y \widehat{P_{Y|X}}(\omega). \end{aligned} \quad (43)$$

Thus, the linear operation first multiplies P_Y by y and then applies the linear shift-invariant transform:

$$\widehat{m}(\omega) = \frac{1}{i \frac{d}{d\omega} \ln(\widehat{P_W}(\omega))}. \quad (44)$$

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. (42) with

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

Substituting into Eq. (44) gives

$$\widehat{m}(\omega) = e^{i\omega}. \quad (45)$$

Substituting this into Eq. (43), taking the inverse Fourier transform and substituting into Eq. (10), gives the estimator

$$\mathbf{E}(X|Y = n) = \frac{(n+1)P_Y(n+1)}{P_Y(n)}, \quad (46)$$

which can be verified by direct calculation.

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) [28]. In this case Eq. (42) holds for

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

In this case, the operator will be

$$\widehat{m}(\omega) = \frac{-1}{i\omega}, \quad (47)$$

which gives

$$\mathbf{E}(X|Y = y) = \frac{-(H(y) - \frac{1}{2}) \star (yP_Y(y))}{P_Y(y)}, \quad (48)$$

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

$$\begin{aligned} \mathbf{E}(X|Y = y) &= \frac{-(H(y)) \star (yP_Y(y))}{P_Y(y)} \\ &= \frac{-\int_{-\infty}^y \tilde{y}P_Y(\tilde{y})d\tilde{y}}{P_Y(y)} \\ &= \frac{-E_Y\{Y; Y < y\}}{P_Y(y)}, \end{aligned} \quad (49)$$

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 .

A third special case occurs when $Y \sim \mathcal{N}(aX, X)$. That is, Y is aX corrupted 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. (42) will hold for

$$\widehat{P}_W(\omega) = e^{-\frac{1}{2}\omega^2 - i a \omega}.$$

In this case, the prior-free operator is

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

where

$$\begin{aligned} \mathcal{F}^{-1}\left\{\frac{1}{a - i\omega}\right\} &= \begin{cases} -e^{ax} I_{\{x>0\}} & a < 0 \\ e^{ax} I_{\{x<0\}} & a > 0 \end{cases} \\ &= \text{sgn}(a) e^{ax} I_{\{ax<0\}} \end{aligned}$$

So that

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

A fourth special case is when X is a random positive value, W is an independent variable drawn from an α -stable distribution [29] 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.(42) 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 prior-free operator is

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

So that

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

B.4 Exponential families

Another important case in which the linear operator may be obtained explicitly is that of the exponential family.

Discrete exponential. First, consider the discrete exponential family of the form

$$Pr\{Y = n|X = x\} = h(x)g(n)x^n, \quad (50)$$

where h is chosen so that summing over n gives one. This case includes the Poisson case discussed in the previous section, amongst others. In this case, we may verify by inspection that

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

so that

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

which tells us that(see [3]):

$$\mathbf{E}(X|Y = n) = \frac{g(n)P_Y(n + 1)}{g(n + 1)P_Y(n)}.$$

Also, we note from Eq. (15) that the linear operator which corresponds to $\frac{1}{X}$ will be

$$\frac{g(n)}{g(n - 1)}P_Y(n - 1) \quad (51)$$

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

$$Pr\{Y = n|X = x\} = h(x)g(n)x^{-n}, \quad (52)$$

we will have

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

Continuous exponential. 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 (this case includes the GSM discussed in the previous section). In this case,

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

So we have

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

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

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

we then have

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

and so

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

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

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

so that

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

and

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

Acknowledgements

This work was partially funded by the Howard Hughes Medical Institute, and by New York University through a McCracken Fellowship to MR.

References

- [1] H. Robbins, “An empirical bayes approach to statistics,” *Proc. Third Berkley Symposium on Mathematical Statistics*, vol. 1, pp. 157–163, 1956.
- [2] K. Miyasawa, “An empirical bayes estimator of the mean of a normal population,” *Bull. Inst. Internat. Statist.*, vol. 38, pp. 181–188, 1961.
- [3] J. S. Maritz and T. Lwin, *Empirical Bayes Methods*, 2nd ed. Chapman & Hall, 1989.
- [4] C. M. Stein, “Estimation of the mean of a multivariate normal distribution,” *Annals of Statistics*, vol. 9, no. 6, pp. 1135–1151, November 1981.
- [5] J. Berger, “Improving on inadmissible estimators in continuous exponential families with applications to simultaneous estimation of gamma scale parameters,” *The Annals of Statistics*, vol. 8, pp. 545–571, 1980.
- [6] J. T. Hwang, “Improving upon standard estimators in discrete exponential families with applications to poisson and negative binomial cases,” *The Annals of Statistics*, vol. 10, pp. 857–867, 1982.

- [7] D. Donoho and I. Johnstone, “Adapting to unknown smoothness via wavelet shrinkage,” *J American Stat Assoc*, vol. 90, no. 432, December 1995.
- [8] A. Benazza-Benyahia and J. C. Pesquet, “Building robust wavelet estimators for multicomponent images using Stein’s principle,” *IEEE Trans. Image Proc.*, vol. 14, no. 11, pp. 1814–1830, November 2005.
- [9] F. Luisier, T. Blu, and M. Unser, “SURE-based wavelet thresholding integrating inter-scale dependencies,” in *Proc IEEE Int’l Conf on Image Proc*, Atlanta GA, USA, October 2006, pp. 1457–1460.
- [10] M. Raphan and E. P. Simoncelli, “Learning to be Bayesian without supervision,” in *Adv. Neural Information Processing Systems 19*, B. Schölkopf, J. Platt, and T. Hofmann, Eds., vol. 19. Cambridge, MA: MIT Press, May 2007, pp. 1145–1152.
- [11] T. Blu and F. Luisier, “The SURE-LET approach to image denoising,” *IEEE Trans. Image Proc.*, vol. 16, no. 11, pp. 2778–2786, November 2007.
- [12] M. Raphan and E. P. Simoncelli, “Optimal denoising in redundant representations,” *IEEE Trans Image Processing*, vol. 17, no. 8, pp. 1342–1352, Aug 2008.
- [13] C. Chau, L. Duval, A. Benazza-Benyahia, and J. Pesquet, “A nonlinear Stein based estimator for multichannel image denoising,” *IEEE Trans. Image Proc.*, vol. 56, no. 8, pp. 3855–3870, Aug 2008.
- [14] A. Hyvärinen, “Estimation of non-normalized statistical models by score matching,” *Journal of Machine Learning Research*, vol. 6, pp. 695–709, 2005.
- [15] A. Hyvärinen, “Optimal approximation of signal priors,” *Neural Computation*, vol. 20, pp. 3087–3110, 2008.
- [16] M. Raphan and E. P. Simoncelli, “Empirical Bayes least squares estimation without an explicit prior,” in *SIAM Conf. on Imaging Science*, Minneapolis, MN, May 2006.
- [17] M. Raphan, “Optimal estimation: Prior free methods and physiological application,” Ph.D. dissertation, Courant Institute of Mathematical Sciences, New York University, New York, NY, May 2007.
- [18] D. Comaniciu and P. Meer, “Mean shift: A robust approach toward feature space analysis,” *IEEE Pat. Anal. Mach. Intell.*, vol. 24, pp. 603–619, 2002.
- [19] G. Casella, “An introduction to empirical Bayes data analysis,” *Amer. Statist.*, vol. 39, pp. 83–87, 1985.
- [20] L. Wasserman, *All of Statistics: A Concise Course in Statistical Inference*. Springer, 2004.
- [21] R. Kalman, “A new approach to linear filtering and prediction problems,” *ASME Transactions*, vol. 82D, 1960.
- [22] M. Raphan and E. P. Simoncelli, “Empirical bayes least squares estimation without an explicit prior,” Computer Science Technical Report, Courant Inst. of Mathematical Sciences, New York University, Tech. Rep. TR2007-900, May 2007.
- [23] D. W. Scott, *Multivariate Density Estimation: Theory, Practice, and Visualization*. John Wiley, 1992.
- [24] C. R. Loader, “Local likelihood density estimation,” *Annals of Statistics*, vol. 24, no. 4, pp. 1602–1618, 1996.

- [25] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdú, and M. Weinberger, “Universal discrete denoising: Known channel,” *IEEE Trans. Info. Theory*, vol. 51, no. 1, pp. 5–28, January 2005.
- [26] W. W. Hager, “Updating the inverse of a matrix,” *SIAM Review*, vol. 31, pp. 221–239, 1989.
- [27] R. M. Wilcox, “Exponential operators and parameter differentiation in quantum physics,” *Journal of Mathematical Physics*, vol. 8, pp. 962–982, 1967.
- [28] D. Andrews and C. Mallows, “Scale mixtures of normal distributions,” *J. Royal Stat. Soc.*, vol. 36, pp. 99–102, 1974.
- [29] W. Feller, *An introduction to probability theory and its applications*. Wiley, 1970, vol. 2.