Computer Science Technical Report TR2007-900, May 2007
Courant Institute of Mathematical Sciences, New York University
http://cs.nyu.edu/web/Research/TechReports/reports.html

Empirical Bayes Least Squares Estimation without an Explicit Prior

Martin Raphan
Courant Inst. of Mathematical Sciences
New York University
raphan@cims.nyu.edu

Eero P. Simoncelli
Howard Hughes Medical Institute,
Center for Neural Science, and
Courant Inst. of Mathematical Sciences
New York University
eero.simoncelli@nyu.edu

Bayesian estimators are commonly constructed using an explicit prior model. In many applications, one does not have such a model, and it is difficult to learn since one does not have access to uncorrupted measurements of the variable being estimated. In many cases however, including the case of contamination with additive Gaussian noise, the Bayesian least squares estimator can be formulated directly in terms of the distribution of *noisy* measurements. We demonstrate the use of this formulation in removing noise from photographic images. We use a local approximation of the noisy measurement distribution by exponentials over adaptively chosen intervals, and derive an estimator from this approximate distribution. We demonstrate through simulations that this adaptive Bayesian estimator performs as well or better than previously published estimators based on simple prior models.

1 Introduction

Denoising is a classic signal processing problem, and Bayesian methods can provide well formulated and highly successful solutions. Bayesian estimators are derived from three fundamental components: 1) the likelihood function, which characterizes the distortion process, 2) the prior, which characterizes the distribution of the variable to be estimated in the absence of any measurement data, and 3) the loss function, which expresses the cost of making errors. Of these, the choice of a prior is often the most difficult aspect of formulating the problem. If the prior is not known in advance, it must be learned from uncorrupted samples(if available), or from noise-corrupted data.

If one uses a least squares loss function, however, it turns out that in many cases the Bayes estimator can be written as a simple expression in terms of the density of *noisy* observations [1, 2, 3, 4]. In this paper we introduce the concept of a "prior-free" Bayesian estimator for the additive Gaussian noise case, develop it into a nonparametric implementation for denoising photographic images, and demonstrate that the denoising results are competitive with those of methods that make use of explicit assumptions or knowledge of the prior.

1.1 Bayes denoising: Conventional formulation

Suppose we make an observation, Y, of a noise-corrupted version of variable X, where either or both variables can be finite-dimensional vectors. Given this observation, we wish to obtain the estimate of X that minimizes the expected squared error. This is a classic problem, and the solution, known as the Bayesian Least Squares (BLS), or Minimum Mean Square Estimate (MMSE), is simply the expected value of X conditioned on Y, $E\{X|Y\}$. If the prior distribution on X is $P_X(\mathbf{x})$ then this can be written using Bayes' rule as

$$\begin{aligned} E\{X|Y=\mathbf{y}\} &= \int \mathbf{x}\, P_{X|Y}(\mathbf{x}|\mathbf{y})\, d\mathbf{x} \\ &= \int \mathbf{x}\, P_{Y|X}(\mathbf{y}|\mathbf{x})\, P_X(\mathbf{x})\, d\mathbf{x} \Big/ P_Y(\mathbf{y})\,, \end{aligned} \tag{1}$$

where the denominator contains the distribution of the noisy observations:

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

Although this approach is generally appealing, it is often criticized for the reliance on knowledge of the prior distribution, $P_X(\mathbf{x})$. In some applications, the prior is known or can be estimated through a set of offline measurements. But in many cases, it must be learned from the same noisy measurements, Y, that are available in the estimation problem. The resulting dilemma presents a problem for machine as well as biological systems: How can a denoiser learn to denoise without having ever seen clean data?

1.2 Bayesian denoising: Prior-free formulation

Surprisingly, under restricted conditions, the BLS estimate may be written without explicit reference to the prior distribution. Specifically, in the case of corruption by additive Gaussian noise, the BLS estimator can be expressed entirely in terms of the distribution of the "noisy" measurements, $P_Y(y)$:

$$E\{X|Y=\mathbf{y}\} = \mathbf{y} + \frac{\Lambda \nabla_{\mathbf{y}} P_Y(\mathbf{y})}{P_Y(\mathbf{y}))} \tag{3}$$

$$= \mathbf{y} + \Lambda \nabla_{\mathbf{y}} ln(P_Y(\mathbf{y})), \tag{4}$$

where Λ is the covariance matrix of the noise.[2] The proof of this fact is straightforward. First, we write the observation equation for additive Gaussian noise contamination:

$$P_{Y|X}(\mathbf{y}|\mathbf{x}) = \frac{1}{(2\pi)^{n/2}|\Lambda|^{1/2}} e^{-\frac{1}{2}(\mathbf{y}-\mathbf{x})^T \Lambda^{-1}(\mathbf{y}-\mathbf{x})} \tag{5}$$

Next, note that

$$\nabla_{\mathbf{y}} P_{Y|X}(\mathbf{y}|\mathbf{x}) = \Lambda^{-1} P_{Y|X}(\mathbf{y}|\mathbf{x})(\mathbf{x}-\mathbf{y})\,. \tag{6}$$

Taking the gradient of

$$P_Y(\mathbf{y}) = \int P_{Y|X}(\mathbf{y}|\mathbf{x}) P_X(\mathbf{x}) d\mathbf{x} \tag{7}$$

with respect to $\mathbf{y}$, dividing by $P_Y(\mathbf{y})$, and substituting Eq. (6) yields:

$$\begin{aligned} \frac{\nabla_{\mathbf{y}} P_Y(\mathbf{y})}{P_Y(\mathbf{y})} &= \frac{\int P_X(\mathbf{x})\, \nabla_{\mathbf{y}} P_{Y|X}(\mathbf{y}|\mathbf{x})\, d\mathbf{x}}{P_Y(\mathbf{y})} \\ &= \frac{\Lambda^{-1} \int P_X(\mathbf{x})\, P_{Y|X}(\mathbf{y}|\mathbf{x})\, (\mathbf{x}-\mathbf{y})\, d\mathbf{x}}{P_Y(\mathbf{y})} \\ &= \Lambda^{-1} \int P_{X|Y}(\mathbf{x}|\mathbf{y})\, (\mathbf{x}-\mathbf{y})\, d\mathbf{x} \\ &= \Lambda^{-1} \left[E\{X|Y=\mathbf{y}\} - \mathbf{y}\right]\,. \end{aligned} \tag{8}$$

Finally, rearranging the terms gives Eq. (3). In what follows, we will restrict ourselves to discussing the case of scalar data.

2 Learning the estimator function from data

The formulation of Eq. (3) offers a means of computing the BLS estimator from noisy samples if one can construct an approximation of the noisy distribution. But simple histograms will not suffice for this approximation, because Eq.(3) require us to compute the logarithmic derivative of the distribution.

2.1 Approximating local logarithmic derivative

A natural solution for this problem is to approximate the logarithmic derivative of the density at the observation $Y_k = y$ as being constant over some interval (x_0, x_1) containing y. This is equivalent to assuming that the density is approximately exponential in the interval:

$$P_Y(y) = ce^{-ay}, \qquad x_0 < y < x_1 \tag{9}$$

where a is the estimate of the logarithmic derivative in the interval (x_0, x_1). Note that it is the a's which are to be used for the estimator, while the c's are irrelevant. For this reason, we look at the conditional density of y given that y is in the interval (x_0, x_1)

$$\begin{aligned} P_{Y|Y\in(x_0,x_1)}(y) &= \frac{e^{-ay}}{\int_{x_0}^{x_1} e^{-ay}dy} I_{(x_0,x_1)} \\ &= \frac{\frac{a}{2}e^{-a(y-\bar{x})}}{\sinh(\frac{a}{2}\Delta x)} I_{(x_0,x_1)} \end{aligned} \tag{10}$$

where $I_{(x_0,x_1)}$ denotes the indicator function of (x_0, x_1), $\bar{x} = \frac{x_0+x_1}{2}$ and $\Delta x = x_1 - x_0$. Comparing this with Eq. (9), we see that the conditional density is also an exponential function of y over the interval (x_0, x_1), with the same exponent a, but is normalized so that c no longer appears, and so that it integrates to one over the interval. If we then have observations Y_n drawn from $P_Y(y)$, and keep only data which fall in (x_0, x_1), these data will have distribution $P_{Y|Y\in(x_0,x_1)}(y)$, so we can use this to estimate the parameter a.

One very popular estimator used for such a problem is the Maximum Likelihood (ML) estimator. Assuming that Eq.(10) is a good approximation of the conditional density on (x_0, x_1), this estimator can be written

$$\begin{aligned} \hat{a} &= \arg\max_a \sum_{\{n:Y_n\in(x_0,x_1)\}} \ln(P_{Y|Y\in(x_0,x_1)}(Y_n)) \\ &= \arg\max_a \{\ln(a) - a(\bar{Y} - \bar{x}) - \ln(\sinh(\frac{a}{2}\Delta x))\} \end{aligned} \tag{11}$$

where

$$\bar{Y} \stackrel{def}{=} \frac{1}{\#\{Y_n \in (x_0, x_1)\}} \sum_{Y_n\in(x_0,x_1)} Y_n \tag{12}$$

is the average of the data that fall into (x_0, x_1). Setting the derivative of Eq. (11) with respect to a equal to zero yields

$$\frac{1}{\hat{a}} - (\bar{Y} - \bar{x}) - \coth(\frac{\hat{a}}{2}\Delta x)\frac{\Delta x}{2} = 0 \tag{13}$$

or

$$\frac{1}{\frac{\hat{a}\Delta x}{2}} - \coth(\frac{\hat{a}\Delta x}{2}) = \frac{\bar{Y} - \bar{x}}{\frac{\Delta x}{2}} \tag{14}$$

Solving this for $\hat{a}$ gives

$$\hat{a} = \frac{2}{\Delta x} f^{-1}\left(\frac{\bar{Y} - \bar{x}}{\frac{\Delta x}{2}}\right) \tag{15}$$

where

$$f(y) = \frac{1}{y} - \coth(y) \tag{16}$$

This local exponential approximation is similar to that used in [5] for local density estimation except that, since we are approximating the local *conditional* density, c disappears from the equation for $\hat{a}$. This has the benefit that we only need to invert a scalar function of one variable, f, to calculate the estimate at all points, instead of inverting a two dimensional vector function of two variables, as is done in [5].

Obviously, it is $\bar{Y}$, the local mean, which requires the most calculation, but, since most of this calculation comes from adding up the value of data which fall in the interval, this may be done in an iterative way, subtracting or adding from a running sum. This method is efficient enough that it may be calculated at each data point, rather than on a grid with interpolation.

2.2 Choice of binwidth

In order to calculate Eq. (15) for a particular y, it is necessary to choose the interval (x_0, x_1), or, equivalently, to choose the binwidth $h = x_1 - x_0$. To define what we mean by an optimal binwidth, we must choose a measure of how "good" an estimate is. We will use the MSE of the estimate, which may be separated into a variance term and a bias term

$$\begin{aligned} E\{(\hat{a} - a)^2\} &= E\{((\hat{a} - E\{\hat{a}\}) + (E\{\hat{a}\} - a))^2\} \\ &= E\{(\hat{a} - E\{\hat{a}\})^2\} + (E\{\hat{a}\} - a)^2 \\ &= Var\{\hat{a}\} + (E\{\hat{a}\} - a))^2 \end{aligned} \tag{17}$$

where $\hat{a}$ is the data-dependent estimate of the true value a. The first term is the variance of the estimator, $\hat{a}$ and will decrease as the binwidth of the interval is increased, since more data will fall into the interval, giving a more reliable estimate. The second term is the squared bias, which will conversely increase as the interval is increased, since the exponential fit of the density over the interval will in general become worse, which means that the estimate $\hat{a}$ will not give a good estimate of the true value of the logarithmic derivative, a. Thus we have a bias-variance tradeoff.

In order to choose an optimal binwidth, we must analyze how Eq. (17) behaves as a function of the binwidth, h. For large amounts of data, we expect h to be small, and so we may use small h approximations for the bias and variance. In general, the variance in estimating the parameter, a, for the interval (x_0, x_1) will depend inversely on the amount of data which falls in the interval. If there are N total data points, we can approximate the number falling in the interval (x_0, x_1) as

$$n \approx P_Y(y) N h \tag{18}$$

Hence, we will assume that

$$Var\{\hat{a}\} \approx \frac{C}{P_Y(y) N h} \tag{19}$$

for an appropriate constant, C.

On the other hand, the squared bias will generally depend only on how well the exponential fits the true density over the interval. As $h \to 0$ the bias for the interval will decrease to zero. For small h we assume that

$$(E\{\hat{a}\} - a)^2 \approx D h^m \tag{20}$$

where $D = D(P_Y, y)$ depends only on the shape of P_Y in the interval, but not on the actual value $P_Y(y)$ (see [5]). In what follows, we will assume that the density is smooth enough that we may ignore the dependence of D on shape, and treat D as constant for all values of y. Since, in our case, P_Y comes from convolving P_X with a Gaussian, P_Y will be at least as smooth as P_X, and will become smoother as the noise variance increases. Therefore, this approximation will become better as the amount of noise increases.

Putting everything together than yields the approximation

$$E\{(\hat{a} - a)^2\} \approx \frac{C}{P_Y(y)Nh} + Dh^m \tag{21}$$

Setting the derivative of this equation with respect to h equal to zero yields

$$Dmh^{m+1} - \frac{C}{P_Y(y)N} = 0 \tag{22}$$

or

$$h = (\frac{C}{DmP_Y(y)N})^{\frac{1}{m+1}} \tag{23}$$

which verifies our assumption that $h \to 0$ as the amount of data increases. Substituting this into Eq. (21) gives

$$E\{(\hat{a} - a)^2\} = \left(\frac{(DmC^m)^{\frac{1}{m+1}}}{(P_Y(y)^{\frac{m}{m+1}})} + D^{\frac{1}{m+1}} (\frac{C}{mP_Y(y)})^{\frac{m}{m+1}} \right) \frac{1}{N^{\frac{m}{m+1}}} \tag{24}$$

which shows that both the squared bias and variance, and hence the MSE, go to zero as $N \to \infty$. Using Eq. (18) to approximate P_Y in Eq. (23) gives

$$h \approx (\frac{Ch}{Dmn})^{\frac{1}{m+1}} \tag{25}$$

Rearranging this equation gives

$$nh^m = \frac{C}{Dm} \tag{26}$$

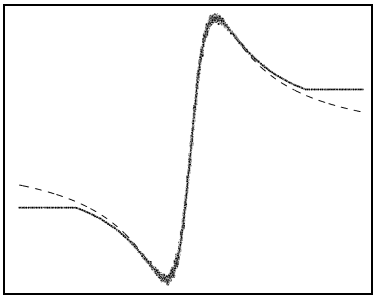
and thus the optimal binwidth is chosen such that the product of the number of points which fall in the interval times some power of the binwidth of the interval is constant. (For a review of bandwidth selection methods for density estimation see [6]. It does not seem that our method of bandwidth selection has been suggested.)

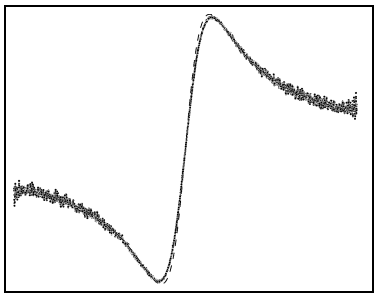
2.3 Choice of power

To determine the binwidth, it is necessary to determine the constant m. If $m = 0$, then n, the number of data points in the neighborhood, will be constant for all data points, a method known as k nearest neighbors (KNN). In the limit as $m \to \infty$, the binwidth will be fixed at a constant value for all data points. Assuming that the approximation of the true logarithmic derivative of the density by a constant is of first order in h leads to the result that the squared bias will be of order h^2, which gives $m = 2$ in Eq. (20). This may be justified by the use of Taylor series when h is very small. In this case there will be an interplay between the binwidth and number of points in the interval.

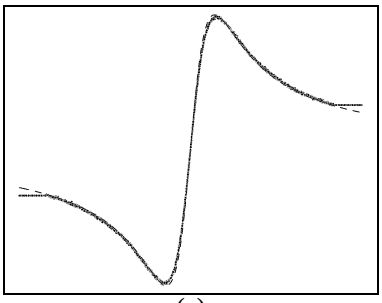
In this section we compare the empirical behavior of binwidths chosen with $m = \{0, 2, \infty\}$, to see how they behave for two different distributions. To put all three methods on the same footing, the constant product for each is chosen so that the average binwidth across data points is the same. Thus, we are looking at how well the three methods allocate this average binwidth.

(a)





(b)



(c)

Fig. 1: Estimate of logarithmic derivative of Cauchy (dashed line is actual value) (**a**) using KNN ($m = 0$); (**b**) using fixed binwidth ($m = \infty$); (**c**) using $m = 2$ to select binwidth

The first density we examine is the Cauchy distribution.

$$P_Y(y) \propto \frac{1}{1 + 0.5y^2} \tag{27}$$

so that

$$\frac{d}{dy} \ln(P_Y(y)) = \frac{y}{1 + 0.5y^2} \tag{28}$$

Figure 1 shows the behavior of the estimate of the logarithmic derivative for the three different methods of binwidth selection for a sample of $9,000$ points drawn from the Cauchy distribution. As can be seen, the KNN method ($m = 0$) has a systematic bias in the tails, the fixed binwidth ($m \to \infty$) method has larger variance in the tails, while the $m = 2$ method has reduced the bias seen in the KNN method without introducing the variance present in the fixed binwidth method.

Now consider the Laplacian distribution

$$P_Y(y) \propto e^{-|x|} \tag{29}$$

which gives

$$\frac{d}{dy} \ln(P_Y(y)) = \mathrm{sgn}(x) \tag{30}$$

Figure 2 shows the behavior of the estimate of the logarithmic derivative for the three different methods of binwidth selection on $9,000$ points drawn from the Laplacian distribution. Notice that in this case, since the logarithmic derivative *is* constant away from the origin, there is no bias problem. As can be seen in this case, the KNN method has more of a variance problem near the origin, the fixed binwidth method has larger variance in the tails, while the $m = 2$ method has reduced the variance near the origin without introducing variance in the tails. Based on these two examples, in what follows we will restrict ourselves to using the $m = 2$ method.

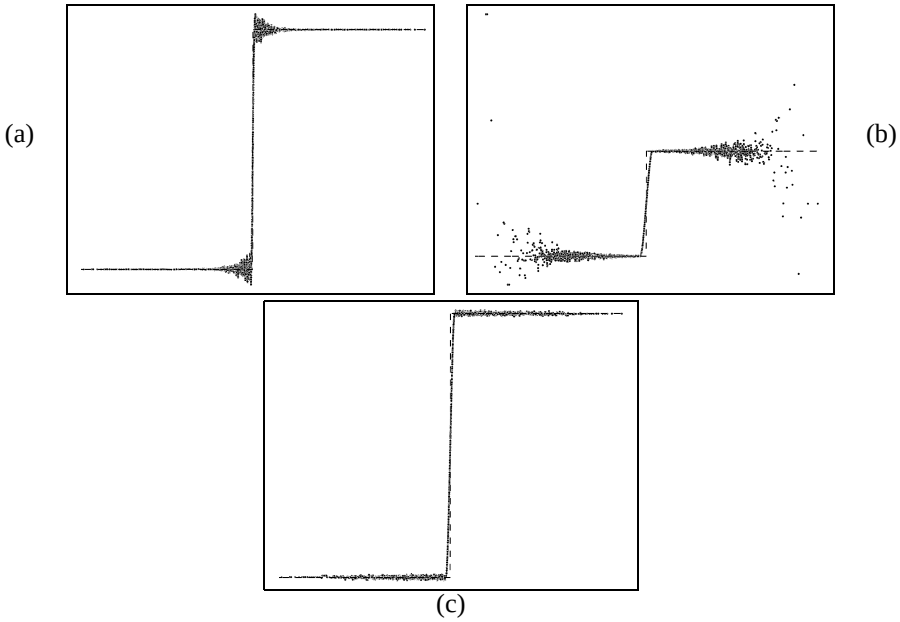


Fig. 2: Estimate of logarithmic derivative of Laplacian (dashed line is actual value) (**a**) using KNN ($m = 0$); (**b**)using fixed binwidth ($m = \infty$; (**c**) using $m = 2$ to select binwidth

The next question is how to choose the average binwidth. Equivalently, we are trying to determine the constant value of the product in Eq. (26). In the examples that follow, we will choose the constant so that the average binwidth across the data is proportional to $\sigma_Y N^{-\frac{1}{m+1}}$, where σ_Y is the standard deviation of the observed data Y. The dependence on σ_Y stems from the intuition that if the data are multiplied by some constant the density will simply be stretched out by that factor, and the binwidth should grow proportionally. The behavior as a function of N comes directly from Eq. (23).

Now that we have a method of binwidth selection, $\bar{Y}$,$\bar{x}$ and Δx, can all be calculated, then Eq. (15) applied to obtain the estimate of the logarithmic derivative, which is then used in Eq. (3) to obtain the BLS estimator.

3 Convergence to ideal BLS estimator with increase in data

Since each bin shrinks and the amount of data in each bin increases with increasing amounts of data, our BLS estimator will approach the ideal BLS estimator as the amount of data increases. In Fig. 3, we illustrate this behavior. For this figure, the density of the prior signal is a generalized Gaussian distribution (GGD)

$$P_X(\mathbf{x}) \propto e^{-|\mathbf{x}/\mathbf{s}|^p} \ . \tag{31}$$

with $s = 1$, and exponent $p = 0.5$. 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, drawing N samples from the prior distribution, corrupting them with additive univariate Gaussian noise, applying the prior free estimator to the data and measuring the resulting SNR. Figure 3 shows the mean improvement in empirical SNR (relative to the ML estimator, which is the identity function), the mean improvement using the conventional

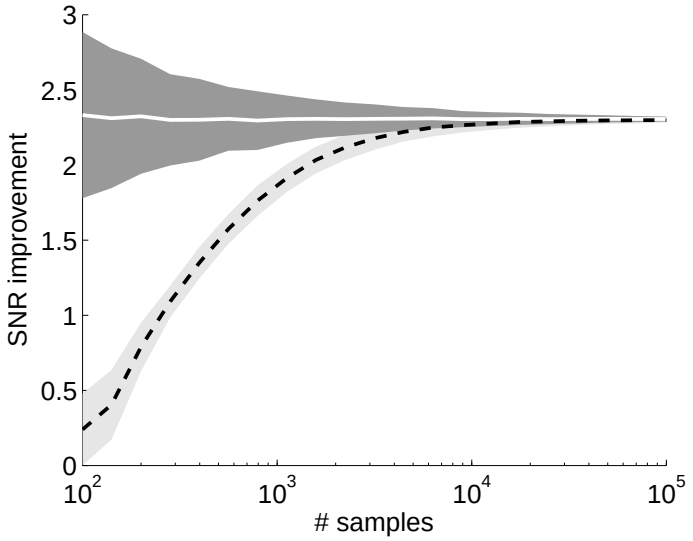


Fig. 3: Empirical convergence of prior-free estimator to optimal BLS solution, as a function number of observed samples of Y. For each number of observations, each estimator is simulated many times. Black dashed lines show the improvement of the prior-free estimator, averaged over simulations, relative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

BLS estimation function,

$$\begin{aligned} E\{X|Y=y\} &= \frac{\int x P_X(x) P_{Y|X}(y|x) dx}{\int P_X(x) P_{Y|X}(y|x) dx} \\ &= \frac{\int x P_X(x) e^{\frac{-(y-x)^2}{2\sigma^2}} dx}{\int P_X(x) e^{\frac{-(y-x)^2}{2\sigma^2}} dx}, \end{aligned} \tag{32}$$

and the standard deviations of these improvements taken over our Monte Carlo simulations.

As can be seen, our estimator improves in performance as it is given more data, and approaches the performance of the ideal BLS estimator as the amount of data increases. It does this without making any assumption about the prior density of the data, instead adapting to the data it does observe. As can also be seen, the variance of this estimator is quite low, for even moderate amounts of data.

4 Comparison with Empirical Bayes

As we have discussed, our prior free estimator will adapt to the observed data, and, given enough data, will give behavior that is near ideal, regardless of the form of the prior distribution. If, instead, we were to assume a particular parametric form for the prior distribution, as in the commonly used Empirical Bayes methods[7], and the true prior did not fall into this parametric family, then the behavior of this estimator would likely be compromised. Thus, our estimator gives a potential advantage over methods which use parametric forms for estimators, since it makes no assumptions about the prior distribution. In exchange, it may require more data than a parametric method. In this section, we will compare the

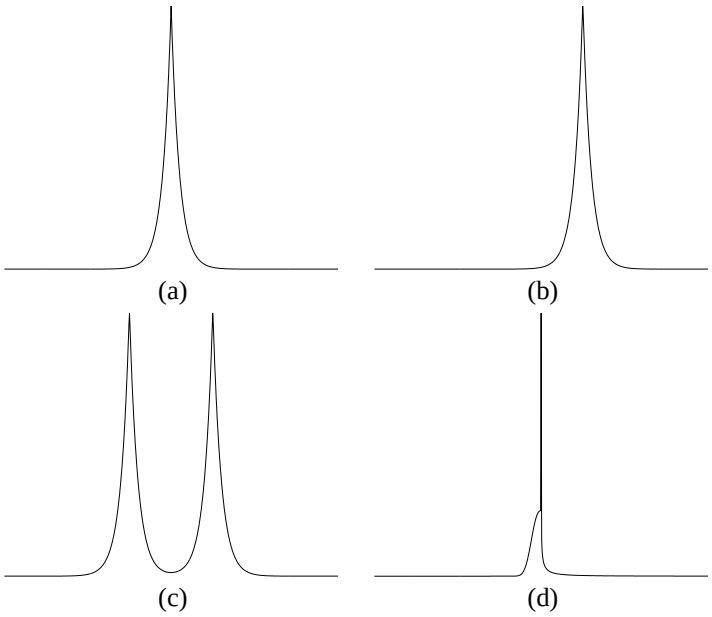


Fig. 4: Other Priors: (**a**) Laplacian (**b**) shifted Laplacian (**c**) bimodal Laplacian (**d**) asymmetric GGD

empirical behavior of our estimator with that of a parametric estimator under conditions where the assumptions of the parametric estimator are valid and under conditions where these assumptions are false.

For our simulations, the Empirical Bayes estimator, based on [8], assumes a GGD form for the prior, as in Eq. (31). The parameters, p and s, are fit to the noisy observation by maximizing the likelihood of the noisy data, and the estimator is computed by numerical integration of

$$\hat{X}_{GGD}(y) = \frac{\int x e^{-|x/s|^p} e^{\frac{-(y-x)^2}{2\sigma^2}} dx}{\int e^{-|x/s|^p} e^{\frac{-(y-x)^2}{2\sigma^2}} dx} \tag{33}$$

and this estimator is then applied to the noisy observations.

4.1 Prior Distributions

The priors we will deal with are shown in Fig. 4. The first is the Laplacian prior (a special case of the GGD), the second is a Laplacian prior with shifted mean, the third is a bimodal Laplacian

$$P_X(x) \propto \frac{1}{2} e^{-|x-m|} + \frac{1}{2} e^{-|x+m|} \tag{34}$$

and the fourth is an asymmetric GGD:

$$P_X(x) \propto \begin{cases} e^{-|\frac{x}{s_1}|^{p_1}}, & x \leq 0 \\ e^{-|\frac{x}{s_2}|^{p_2}}, & x > 0 \end{cases} \tag{35}$$

where the constants are chosen such that the distribution still has zero mean. Thus, the first distribution fits the model assumed by the Empirical Bayes method, whereas the last three break it in some simple ways.

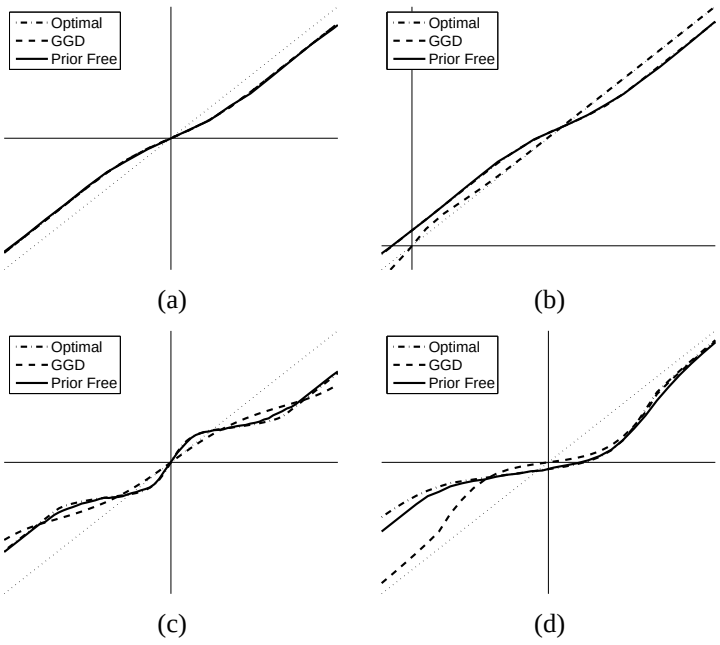


Fig. 5: Coring Functions for: (**a**) Laplacian (**b**) shifted Laplacian (**c**) bimodal Laplacian (**d**) asymmetric GGD prior distributions. In all figures, the dotted line denotes the identity function for reference.

4.2 Results

In these cases, since the prior is known the optimal solution may be calculated directly numerically integrating Eq. (32). Figure 5 shows the estimators, also known as coring functions, obtained for the prior-free and GGD methods from the observed data, as compared with the optimal solution calculated by numerical integration of Eq. (32). Table 4.2 shows the empirical SNR obtained from applying these methods to the observed data, for the priors discussed, as simulated for various values of noise power. Since the eventual application we have in mind is in image processing, we picked $9,000$ data points in our simulation, a reasonable number for such applications.

As is to be expected, in the case where the prior actually fits the assumptions of the GGD model, then the GGD method will outperform the prior-free method, though, it should be noted, not by very much. In the cases where the assumption on the prior is broken in some simple ways, however, the performance of the GGD method degrades considerably while that of the the prior-free method remains surprisingly close to ideal.

5 Image denoising example

In this section we describe a specific example of this prior-free approach as applied to image denoising. The development of multi-scale (wavelet) representations has led to substantial improvements in many signal processing applications, especially denoising. Typically, the signal (or image) is decomposed into frequency bands at multiple scales, each of which is independently denoised by applying a pointwise nonlinear shrinkage function that suppresses low-amplitude values. The concept was developed originally in the television engineering literature (where it is known as “coring”[e.g. 9, 10]), and specific shrinkage

Prior Distn.	Noise SNR	Denoised SNR		
		Opt.	GGD	Prior-free
Lapl.	1.800	4.226	4.225	4.218
	4.800	6.298	6.297	6.291
	7.800	8.667	8.667	8.666
	10.800	11.301	11.301	11.299
Shifted	1.800	4.219	2.049	4.209
	4.800	6.273	4.920	6.268
	7.800	8.655	7.762	8.651
	10.800	11.285	10.735	11.284
Bimodal	1.800	4.572	4.375	4.547
	4.800	7.491	6.767	7.468
	7.800	10.927	9.262	10.885
	10.800	13.651	11.776	13.603
Asym.	1.800	7.102	6.398	7.055
	4.800	8.944	8.170	8.915
	7.800	10.787	10.044	10.767
	10.800	12.811	12.143	12.791

Table 1: Simulated denoising results.

functions have been derived under a variety of formulations, including minimax optimality under a smoothness condition [11, 12, 13], and Bayesian estimation with non-Gaussian priors [e.g. 8, 14, 15, 16, 17, 18, 19, 20]. Note that, although such methods denoise each coefficient separately, a process which will not generally be optimal unless the coefficients are independent (which is impossible for redundant transformations), such marginal denoising methods have proven effective.

As in [8, 17, 21], we begin by decomposing the noisy image using a steerable pyramid. This is a redundant, invertible linear transform that separates the image content into oriented octave-bandwidth frequency subbands. We apply our prior free estimator to each subband separately, using the noisy data in a subband to construct an estimator for that subband. We then apply the subband estimator to the noisy coefficients in the subband in order to estimate the values of the original, noise-free subband. After the coefficients of each subband have been processed, the inverse pyramid transform is applied in order to reconstruct the denoised image.

5.1 Results

We have applied our prior-free Bayesian estimator to several images contaminated with simulated Gaussian noise. For all examples, the noise variance was assumed to be known. The results were compared with two other methods of denoising. The first method [8], described in the last section, uses ML to fit the parameters of a GGD prior, Eq. (31), to the noisy data in the subband. This is justified by the fact that the GGD is a parametric form which is known to provide good fits for the marginal densities of coefficients in image subbands [22, 8, 17, 18]. We then use use this parametric prior to find the associated estimator by numerical integration of Eq. (33).

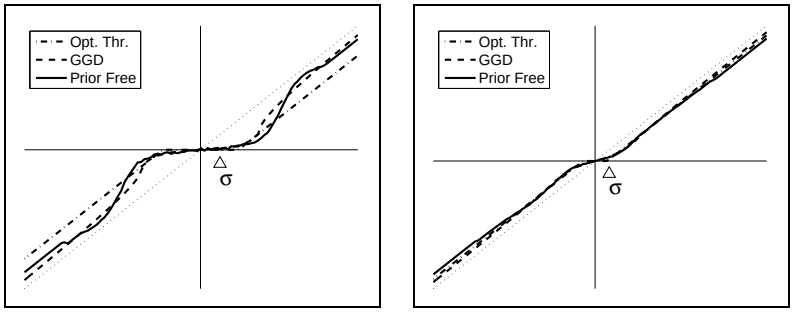


Fig. 6: Example estimators (coring functions) for two subbands: Prior-free Bayesian estimator (solid), BLS estimator for a GGD (dashed), and optimal soft threshold (dash-dotted). Dotted line indicates the identity function. Noise standard deviation σ is also indicated.

The second estimator is a "soft threshold" function, as used in[11]:

$$\hat{x}(Y) = \begin{cases} Y - t, & t \leq Y \\ 0, & -t < Y < t \\ Y + t, & Y \leq -t \, . \end{cases} \tag{36}$$

We make use of the clean, original data to find a soft threshold for each subband that minimizes the empirical mean squared error in that subband. Thus, the performance of this method should not be interpreted as resulting from a feasible denoising algorithm, but rather as an upper bound on thresholding approaches to denoising. Two example estimators are shown in Fig. 6.

Figure 7 shows a sample of an image denoised using these three methods. Table 5.1 shows denoising results for some sample images under several noise conditions. As can be seen, the prior-free approach compares favorably to the other two, despite the fact that it makes weaker assumptions about the prior than does the generalized Gaussian, and doesn't have access to the clean data, as does the optimum thresholding. Figure 8 shows a histogram of PSNR improvement of the prior-free algorithm over optimal thresholding and generalized Gaussian approaches for nine images at four different noise levels. As we can see, our prior free method compares favorably with the parametric method, which was based on detailed empirical knowledge of the statistics of image coefficients.

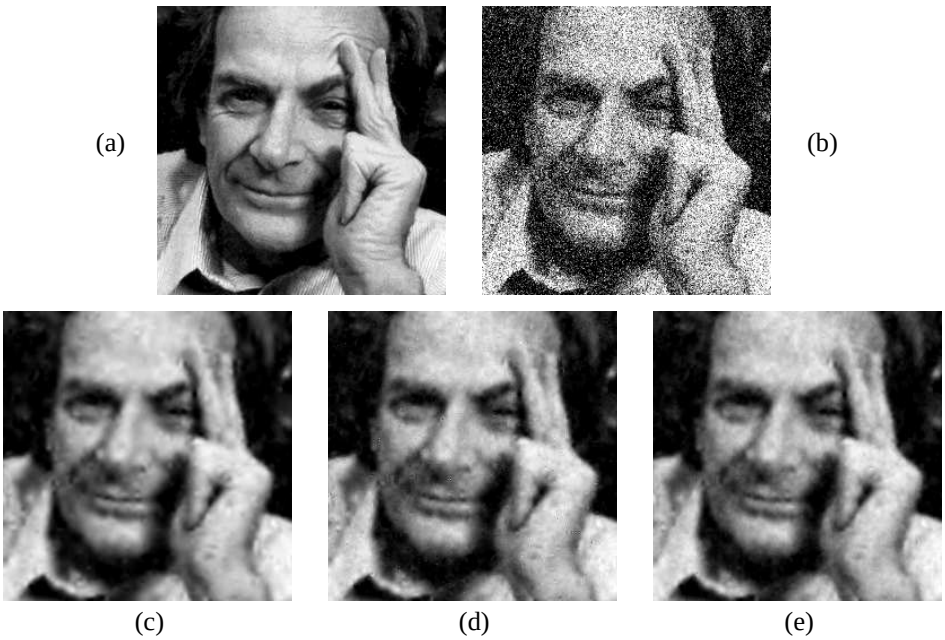


Fig. 7: Denoising results for the “Feynman” image. (**a**) original; (**b**) noisy image (PSNR = 12.71 dB); (**c**) using optimal thresholding (PSNR = 25.02 dB) (**d**) using generalized Gaussian (PSNR =24.77 dB) (**e**) using prior-free denoising (PSNR = 24.86 dB)

Image	Noise	Denoised PSNR		
	PSNR	Opt. Thr.	GGD	Prior-free
crowd	15.8783	26.4656	26.2465	26.333
	18.8783	28.0198	27.8368	27.8779
	21.8783	29.7355	29.5498	29.6008
	24.8783	31.5095	31.37	31.3928
feynman	12.7117	25.0311	24.7549	24.8574
	15.7117	26.1558	26.051	26.0729
	18.7117	27.4194	27.3848	27.3534
	21.7117	28.7006	28.7162	28.6775
boats	16.4778	27.1993	27.0371	27.1585
	19.4778	28.6465	28.5733	28.6439
	22.4778	30.2497	30.2161	30.2799
	25.4778	31.9319	31.9379	32.0262
einstein	17.5359	26.6842	26.5818	26.5132
	20.5359	28.0678	28.0155	27.955
	23.5359	29.4865	29.4828	29.4252
	26.5359	31.0617	31.1044	31.0636
lena	16.3128	28.0438	27.7634	27.8942
	19.3128	29.513	29.3355	29.3937
	22.3128	30.9951	30.8883	30.9357
	25.3128	32.6389	32.5864	32.6361
bench	13.9423	20.1491	20.2218	20.1812
	16.9423	21.4907	21.5634	21.5328
	19.9423	23.1416	23.1816	23.1636
	22.9423	25.0898	25.1185	25.1158
brick	16.5785	22.5231	22.4509	22.421
	19.5785	24.0482	24.0604	24.0453
	22.5785	25.6705	25.7991	25.7848
	25.5785	27.5393	27.7055	27.6933
bridge	15.4273	23.1067	23.0942	23.0978
	18.4273	24.3743	24.3872	24.3676
	21.4273	25.7939	25.8256	25.8115
	24.4273	27.4841	27.5448	27.5285

Table 2: Simulated denoising results.

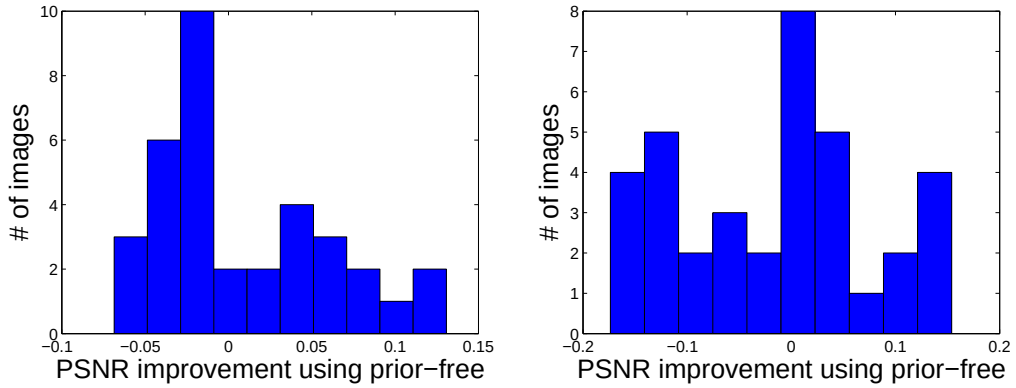


Fig. 8: Improvement in PSNR for prior-free approach compared with the GGD estimator (left) and optimal thresholding (right). Histograms summarize data for 9 images at 4 noise levels.

6 Discussion

We've discussed a modified formulation of the Bayes least squares estimator in the case of additive Gaussian noise. Unlike the traditional form, this estimator is written in terms of the distribution of the noisy measurement data, and is thus more natural for situations in which the prior must be learned from the data. We've developed a local approximation to this prior free formulation, which uses adaptive binwidths to give improved performance with an increase in the number of samples drawn from the noisy distribution. We've shown that as the amount of data is increased, the prior free estimator will tend to give performance that is near ideal. We've also shown that breaking the assumptions of parametric models of the prior leads to a drastic reduction in the performance of methods based on such assumptions, while the prior-free method is able to deal with such changes. Finally, we've demonstrated the feasibility of this methodology by applying it to the problem of image denoising, demonstrating that it performs as well or better than estimators based on marginal prior models found in the literature, which are based on empirical studies of the marginal statistics of clean image subbands. Therefore, in situations where the prior distribution of the clean data is unknown, our method can be used, with some confidence that not too much is lost by not examining and modeling the empirical statistics of clean data, which may not even be possible in some situations.

It must be pointed out that the prior-free method requires a lot of data to be feasible. Also, in cases where an accurate model of the prior is available, methods that make use of this explicit model may give some improvement. However, if nothing is known about the prior, and there is a lot of data, then the prior-free method should give improvement over an ad-hoc assumption about the prior.

In order to obtain image denoising results which are competitive with the state of the art, it is necessary to jointly denoise vectors of coefficients, instead of one coefficient at a time [23, 21]. While Eq. (3) holds for vectors as well as scalars, finding neighborhoods of vectors to use in estimating the logarithmic gradient at a point becomes much more difficult. For higher dimensions the data vectors will tend to be further and further apart (the "curse" of dimensionality), so great care must be taken in choosing the shape of the large neighborhoods required to include sufficient number of data points.

Acknowledgments

This work was partially funded by 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 Mathematcal 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*. Chapman & Hall, 2nd ed., 1989.

[4] M. Raphan and E. P. Simoncelli, "Learning to be Bayesian without supervision," in *Adv. Neural Information Processing Systems (NIPS*06)* (B. Schölkopf, J. Platt, and T. Hofmann, eds.), vol. 19, (Cambridge, MA), MIT Press, May 2007.

[5] C. R. Loader, "Local likelihood density estimation," *Annals of Statistics*, vol. 24, no. 4, pp. 1602–1618, 1996.

[6] D. W. Scott, *Multivariate Density Estimation: Theory, Practice, and Visualization*. John Wiley, 1992.

[7] G. Casella, "An introduction to empirical Bayes data analysis," *Amer. Statist.*, vol. 39, pp. 83–87, 1985.

[8] E. P. Simoncelli and E. H. Adelson, "Noise removal via Bayesian wavelet coring," in *Proc 3rd IEEE Int'l Conf on Image Proc*, vol. I, (Lausanne), pp. 379–382, IEEE Sig Proc Society, September 16-19 1996.

[9] J. P. Rossi, "Digital techniques for reducing television noise," *JSMPTE*, vol. 87, pp. 134–140, 1978.

[10] B. E. Bayer and P. G. Powell, "A method for the digital enhancement of unsharp, grainy photographic images," *Adv in Computer Vision and Im Proc*, vol. 2, pp. 31–88, 1986.

[11] D. L. Donoho and I. M. Johnstone, "Ideal spatial adaptation via wavelet shrinkage," *Biometrika*, vol. 81, pp. 425–455, 1994.

[12] A. Chambolle, R. A. DeVore, and B. J. L. N. Lee, "Nonlinear wavelet image processing: Variational problems, compression, and noise removal through wavelet shrinkage," *IEEE Trans. Image Proc.*, vol. 7, pp. 319–335, March 1998.

[13] D. Leporini and J. C. Pesquet, "Multiscale regularization in Besov spaces," in *31st Asilomar Conf on Signals, Systems and Computers*, (Pacific Grove, CA), November 1998.

[14] H. A. Chipman, E. D. Kolaczyk, and R. M. McCulloch, "Adaptive Bayesian wavelet shrinkage," *J American Statistical Assoc*, vol. 92, no. 440, pp. 1413–1421, 1997.

[15] F. Abramovich, T. Sapatinas, and B. W. Silverman, "Wavelet thresholding via a Bayesian approach," *J R Stat Soc B*, vol. 60, pp. 725–749, 1998.

[16] B. Vidakovic, "Nonlinear wavelet shrinkage with Bayes rules and Bayes factors," *Journal of the American Statistical Association*, vol. 93, pp. 173–179, 1998.

[17] E. P. Simoncelli, "Bayesian denoising of visual images in the wavelet domain," in *Bayesian Inference in Wavelet Based Models* (P. Müller and B. Vidakovic, eds.), ch. 18, pp. 291–308, New York: Springer-Verlag, 1999. Lecture Notes in Statistics, vol. 141.

[18] P. Moulin and J. Liu, "Analysis of multiresolution image denoising schemes using a generalized Gaussian and complexity priors," *IEEE Trans. Info. Theory*, vol. 45, pp. 909–919, 1999.

[19] Hyvarinen, "Sparse code shrinkage: Denoising of nongaussian data by maximum likelihood estimation," *Neural Computation*, vol. 11, no. 7, pp. 1739–1768, 1999.

[20] J. Starck, E. J. Candes, and D. L. Donoho, "The curvelet transform for image denoising," *IEEE Trans. Image Proc.*, vol. 11, pp. 670–684, June 2002.

[21] J. Portilla, V. Strela, M. Wainwright, and E. P. Simoncelli, "Image denoising using a scale mixture of Gaussians in the wavelet domain," *IEEE Trans Image Processing*, vol. 12, pp. 1338–1351, November 2003.

[22] S. G. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Pat. Anal. Mach. Intell.*, vol. 11, pp. 674–693, July 1989.

[23] L. Şendur and I. W. Selesnick, "Bivariate shrinkage functions for wavelet-based denoising exploiting interscale dependency," *IEEE Trans. Sig. Proc.*, vol. 50, pp. 2744–2756, November 2002.