Wavelet-Based Poisson Rate Estimation Using the Skellam Distribution

Keigo Hirakawa[a], Farhan Baqai[b], and Patrick J. Wolfe[a]

[a] School of Engineering and Applied Sciences, Harvard University
33 Oxford Street, Cambridge, MA 02138 USA;
[b] Media Processing Technology Laboratory, Sony Electronics, Inc.
1730 N. First Street, San Jose, CA 95112

Abstract

Owing to the stochastic nature of discrete processes such as photon counts in imaging, real-world data measurements often exhibit heteroscedastic behavior. In particular, time series components and other measurements may frequently be assumed to be non-iid Poisson random variables, whose rate parameter is proportional to the underlying signal of interest—witness literature in digital communications, signal processing, astronomy, and magnetic resonance imaging applications. In this work, we show that certain wavelet and filterbank transform coefficients corresponding to vector-valued measurements of this type are distributed as sums and differences of independent Poisson counts, taking the so-called *Skellam distribution.* While exact estimates rarely admit analytical forms, we present Skellam mean estimators under both frequentist and Bayes models, as well as computationally efficient approximations and shrinkage rules, that may be interpreted as Poisson rate estimation method performed in certain wavelet/filterbank transform domains. This indicates a promising potential approach for denoising of Poisson counts in the above-mentioned applications.

1. INTRODUCTION

Real-world sensing devices are subject various types of measurement noise. For example, it is well known that the lack of resolution (e.g. quantization), randomness inherent in the signal (e.g. photon/packet arrival), and variabilities in the measuring devices (e.g. thermal noise, electron leakage) contribute to a significant degradation of a signal. Estimation of vector-valued data $\boldsymbol{f} \in \mathbb{R}^N$ given noisy observations $\boldsymbol{g} \in \mathbb{R}^N$ therefore plays a prominent role in digital communications, signal processing, astronomy, and magnetic resonance imaging applications.

To illustrate the difficulties associated with estimating the true underlying signal, suppose we adopt a Bayesian statistics point of view; that is, we model the signals in terms of the prior probability distribution of the latent variable ($p(\boldsymbol{f})$) and the likelihood of the observation conditioned on the latent variable ($p(\boldsymbol{g}|\boldsymbol{f})$). Bayesian statistical estimation and inference techniques make use of the posterior probability, or the probability of the latent variable conditioned on the observation ($p(\boldsymbol{f}|\boldsymbol{g})$), which is proportional to the product of the prior probability distribution of the latent variable and the likelihood function. Motivated by the prior knowledge and empirical studies, statistical modeling of the latent variables in the linear transform domains has enjoyed tremendous popularity—in particular, filterbank, wavelets, and short-time Fourier transforms provide convenient platforms for specifying the prior because their coefficients exhibit temporal and spectral locality, sparsity, and energy compaction properties.[1–5]

In this paradigm, the special case of additive white Gaussian noise (AWGN) is studied almost exclusively because the posterior of the transform coefficients is readily accessible when the likelihood function has a closed form in the transform domain. The assumption that the noise is AWGN, however, is inadequate for many real-world applications because the measurement noise is almost always dependent on the range space of the signal $\boldsymbol{f}$, effects of which permeate across multiple transform coefficients and subbands. For instance, the number

Further author information:
K.H.: E-mail: hirakawa@stat.harvard.edu
F.B.: E-mail: farhan.baqai@am.sony.com
P.J.W.: E-mail: patrick@seas.harvard.edu

of electrons or photons encountered in a measuring device during an integration period is a Poisson process $g_i|\boldsymbol{f} \overset{\text{iid}}{\sim} \mathcal{P}(f_i)$ where f_i is the rate parameter (the expected electron/photon count per integration period), and is proportional to the electric current or the light intensity.[6]

One existing strategy to estimating Poisson rate is to design an invertible nonlinear operator $\gamma : \mathbb{R}^N \to \mathbb{R}^N$ that (approximately) maps the heteroscedastic process to a familiar homoscedastic process:[7–9]

$$\gamma(\boldsymbol{g})|\gamma(\boldsymbol{f}) \sim \mathcal{N}(\gamma(\boldsymbol{f}), \boldsymbol{I}).$$

An AWGN-based signal estimation technique is used to estimate $\gamma(\boldsymbol{f})$ given $\gamma(\boldsymbol{g})$, and the inverse transform $\gamma^{-1}(\cdot)$ yields an estimate of $\boldsymbol{f}$. Although this this approach, commonly referred to as *variance stabilization*, modularizes the designs of $\gamma(\cdot)$ and the estimator, the signal model assumed for $\boldsymbol{f}$ no longer holds true for $\gamma(\boldsymbol{f})$ and the optimality of the estimator in the new domain does not translate well to optimality in the range space of $\boldsymbol{f}$. Alternatively, the Poisson distribution $g_i|\boldsymbol{f} \sim \mathcal{P}(f_i)$ tends toward a Normal distribution $\mathcal{N}(f_i, f_i)$ as the integration period increases. Leveraging existing AGWN paradigm once again, we instead consider the problem of the form $g_i|\boldsymbol{f} \sim \mathcal{N}(f_i, \hat{f}_i)$, where $\hat{\boldsymbol{f}}$—a (crude) estimate for $\boldsymbol{f}$—is assumed known and independent of $\boldsymbol{f}$.[10] This approach is not robust to assumptions, and noise variance estimation employs heuristics. In a companion manuscript in preparation, we address variance-stabilizing transforms and other such approximations in detail.

In this paper, we derive a representation of vector-valued Poisson counts in the Haar wavelet (HWT) and Haar filterbank (HFT) transform domains based on the *Skellam distribution*[11]—a distribution whose applications to-date has been limited largely to modeling a difference of Poisson counts[12, 13] or a gradient of Poisson corrupted signal.[14, 15] Though there exist other filterbank/wavelet transforms with better frequency separation, the advantage to encoding the likelihood function in the transform domain is that the posterior distribution of the latent variables is readily accessible. In light of this, we propose two strategies for estimating the Poisson rate from the noise-corrupted signal in the transform domain. First, the structure of noise in the transform domain admits a variant of Stein's unbiased risk estimator (SURE) for a transform-based parametric estimator.[16–18] Second, we design a Bayesian estimator based on a canonical heavy-tailed (i.e., super-Gaussian) prior probability distribution. Although the posterior mean estimator of the Poisson rate does not appear to admit an exact analytical expression, we derive an analytical approximation to the optimal estimator that we show to be both efficient and practical.

The remainder of this paper is organized as follows. We briefly review SURE shrink[19] in Section 2. We then draw connections between Skellam distribution and HWT/HFT in Section 3, prove basic properties about the distribution, and derive the unbiased estimate of risk. In section 4 we propose a Bayesian form of Skellam mean estimator based on Laplacian prior and its analytical approximation. The appendix sections then detail the proofs and mathematical properties exploited in the main body of the paper.

2. BACKGROUND

Suppose for a moment that $\tilde{\boldsymbol{f}} \in \mathbb{R}^N$ and

$$\tilde{\boldsymbol{g}} \sim \mathcal{N}(\tilde{\boldsymbol{f}}, \tilde{\sigma}^2 \boldsymbol{I}).$$

Let $\boldsymbol{W} \in \mathbb{R}^{N \times N}$ be an orthogonal transform matrix, and where $\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}} \in \mathbb{R}^N$, $\tilde{\boldsymbol{x}} = \boldsymbol{W}\tilde{\boldsymbol{f}}$ and $\tilde{\boldsymbol{y}} = \boldsymbol{W}\tilde{\boldsymbol{g}}$ are clean and noisy transform coefficients. Then the orthogonality of transformation $\boldsymbol{W}$ guarantees the following:

$$\tilde{\boldsymbol{y}} \sim \mathcal{N}(\tilde{\boldsymbol{x}}, \tilde{\sigma}^2 \boldsymbol{I}).$$

The estimator for $\tilde{\boldsymbol{x}}$ given $\tilde{\boldsymbol{y}}$ can be written as $\hat{\tilde{\boldsymbol{x}}}(\tilde{\boldsymbol{y}}) = \tilde{\boldsymbol{y}} + \phi(\tilde{\boldsymbol{y}})$. The unbiased estimate of risk, then, is:

$$\begin{aligned}
\mathbb{E}\,\|\hat{\tilde{\boldsymbol{x}}}(\tilde{\boldsymbol{y}}) - \tilde{\boldsymbol{x}}\|^2 =& \mathbb{E}\,\|(\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{x}}) + \phi(\tilde{\boldsymbol{y}})\|^2 \\
=& \mathbb{E}(\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{x}})^2 + 2\,\mathbb{E}(\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{x}})\phi(\tilde{\boldsymbol{y}}) + \mathbb{E}\,\phi(\tilde{\boldsymbol{y}})^2 \\
=& \tilde{\sigma}^2 + 2\tilde{\sigma}^2\,\mathbb{E}\,\triangledown_{\tilde{y}}\phi(\tilde{\boldsymbol{y}}) + \mathbb{E}\,\phi(\tilde{\boldsymbol{y}})^2,
\end{aligned}$$

where the last equality is a result of Stein's lemma,[16–18]

LEMMA 2.1. *(Stein's Lemma) Let $\tilde{y}|\tilde{x} \sim \mathcal{N}(\tilde{x},\sigma^2)$. Then for any function $\phi : \mathbb{R} \to \mathbb{R}$,*

$$\mathbb{E}(\tilde{y}-\tilde{x})\phi(\tilde{\boldsymbol{y}}) = \sigma^2\,\mathbb{E}\,\bigtriangledown_{\tilde{\boldsymbol{y}}}\phi(\tilde{\boldsymbol{y}})$$

This is particularly useful for choosing parameters for parametric estimators—especially when the smoothness of the underlying signal is unknown *a priori*—as the expectation can be replaced by the ensemble of the observed noisy coefficients. For example, soft thresholding with a parameter τ is defined as:[19]

$$\hat{\tilde{x}}_i(\tilde{y}_i) = \operatorname{sgn}(\tilde{y}_i)\big(|\tilde{y}_i| - \tau\big)_+,$$

where $\hat{\tilde{\boldsymbol{x}}}(\tilde{\boldsymbol{y}}) = \tilde{\boldsymbol{y}} + \phi(\tilde{\boldsymbol{y}})$ and

$$\phi(\tilde{y}_i) = \begin{cases} -\operatorname{sgn}(\tilde{y}_i)\tau & \text{if } |\tilde{y}_i| \geq \tau \\ -\tilde{y}_i & \text{if } |\tilde{y}_i| < \tau \end{cases}, \qquad \frac{d}{d\tilde{y}_i}\phi(\tilde{y}_i) = \begin{cases} 0 & |\tilde{y}_i| \geq \tau \\ -1 & \text{if } |\tilde{y}_i| < \tau \end{cases}$$

Thus the minimizer of the expected risk

$$\begin{aligned}\mathbb{E}\,\|\hat{\tilde{x}}(\tilde{\boldsymbol{y}}) - \tilde{\boldsymbol{x}}\|^2 \approx& N\tilde{\sigma}^2 - \sum_{|\tilde{y}_i|<\tau} 2\tilde{\sigma}^2 + \sum_{|\tilde{y}_i|<\tau} \tilde{y}_i^2 + \sum_{|\tilde{y}_i|\geq\tau} \tau^2 \\ =& N\tilde{\sigma}^2 - 2\tilde{\sigma}^2(\#\text{ of } |\tilde{y}_i| < \tau) + \sum_i \min(\tilde{y}_i^2,\tau^2) \end{aligned} \tag{1}$$

is said to be the "best" choice for τ.

3. SKELLAM DISTRIBUTION AND HAAR WAVELET/FILTERBANK

The Haar wavelet and filterbank transforms are attractive alternatives to the more computationally expensive joint time-frequency analysis techniques that yield better frequency localization. They enjoy orthonormality, compact spatial support, and computational simplicity—and the HWT satisfies the axioms of multiresolution analysis. In this section, we demonstrate that the simplicity of these transforms admits analytical tractability that enables inference and estimation.

Assume $g_j|\boldsymbol{i} \overset{\text{iid}}{\sim} \mathcal{P}(f_i)$ as before. Let $\boldsymbol{W} \in \{0,\pm1\}^{N\times N}$ be the forward HWT or HFT matrix*, and $\boldsymbol{x} = \boldsymbol{W}\boldsymbol{f}$ and $\boldsymbol{y} = \boldsymbol{W}\boldsymbol{g}$—in this case, $\boldsymbol{x},\boldsymbol{y}$ are sums and differences of signal sample values in $\boldsymbol{f},\boldsymbol{g}$, respectively. Define x_i^+, x_i^- as positive and negative combinations of $\boldsymbol{f}$ that comprise x_i—that is, $x_i = x_i^+ - x_i^-$, where

$$x_i^+ = \sum_{j:[\boldsymbol{W}]_{i,j}=1} f_j, \qquad x_i^- = \sum_{j:[\boldsymbol{W}]_{i,j}=-1} f_j, \tag{2}$$

and $y_i = y_i^+ - y_i^-$ defined similarly based on $\boldsymbol{g}$. Owing to the fact that a sum of independent Poisson counts constitute another Poisson process, y_i^+ and y_i^- exhibit the following behavior:

$$y_i^+|\boldsymbol{f} \overset{\text{iid}}{\sim} \mathcal{P}(x_i^+), \qquad y_i^-|\boldsymbol{f} \overset{\text{iid}}{\sim} \mathcal{P}(x_i^-).$$

Thus, the transform coefficient $y_i = y_i^+ - y_i^-$ is a difference of two independent Poisson counts, termed the Skellam distribution,[11] parameterized by x_i^+ and x_i^-:

$$\begin{aligned}\forall k \in \mathbb{Z}, \quad p(y_i = k; x_i^+, x_i^-) =& e^{-(x_i^+ + x_i^-)} \sum_{n\geq\min(0,-k)} \frac{(x_i^+)^{k+n}(x_i^-)^n}{(k+n)!n!} \\ =& e^{-(x_i^+ + x_i^-)} \left(\frac{x_i^+}{x_i^-}\right)^{\frac{k}{2}} I_k\left(2\sqrt{x_i^+ x_i^-}\right)\end{aligned}$$

[*] With appropriate normalization of rows, $\boldsymbol{W}$ becomes a unitary matrix. However, the connections to Skellam distribution will be made clearer when $[\boldsymbol{W}]_{ij} \in \{0,\pm1\}$, in which case orthogonality of rows/columns is still retained.

where $I_k(\cdot)$ is the modified Bessel function of the first kind.[20] Its range is the integers $\mathbb{Z}$.

Define $s_i = x_i^+ + x_i^-$ and $t_i = y_i^+ + y_i^-$ as the corresponding scaling coefficients of $\boldsymbol{Wf}$ and $\boldsymbol{Wg}$, respectively, where $t_i|\boldsymbol{f} \overset{\text{iid}}{\sim} \mathcal{P}(s_i)$. When $\boldsymbol{W}$ is the forward HFT matrix, in fact, we can show that the density of an observed transform coefficient y_i is parameterized by x_i and s_i *only.*

Proposition 3.1. *Let $g_j|\boldsymbol{f} \overset{iid}{\sim} \mathcal{P}(f_j)$, $\boldsymbol{x} = \boldsymbol{Wf}$, $\boldsymbol{y} = \boldsymbol{Wg}$ as before. Then*

$$p(y_i|\boldsymbol{f}) = p(y_i|s_i, x_i).$$

Proof. The proof is a straightforward consequence of the choice of transform. From the definitions in (2),

$$\begin{aligned}
p(y_i|\boldsymbol{f}) =& p(y_i|x_i^+, x_i^-) \\
=& p\left(y_i \middle| \sum_{[W]_{i,j}=1} f_j, \sum_{[W]_{i,j}=-1} f_j\right) \\
=& p\left(y_i \middle| \sum_{[W]_{i,j}=1} [\boldsymbol{W}^{-1}\boldsymbol{x}]_j, \sum_{[W]_{i,j}=-1} [\boldsymbol{W}^{-1}\boldsymbol{x}]_j\right).
\end{aligned}$$

Define $\boldsymbol{v}_i$ and $\boldsymbol{w}_i$ as the row vectors from $\boldsymbol{W}$ such that $s_i = \boldsymbol{v}_i\boldsymbol{f}$ and $x_i = \boldsymbol{w}_i\boldsymbol{f}$. Then,

$$\begin{aligned}
p(y_i|\boldsymbol{f}) =& p\left(y_i \middle| \left(\frac{\boldsymbol{v}_i + \boldsymbol{w}_i}{2}\right)\boldsymbol{W}^{-1}\boldsymbol{x}, \left(\frac{\boldsymbol{v}_i - \boldsymbol{w}_i}{2}\right)\boldsymbol{W}^{-1}\boldsymbol{x}\right) \\
=& p\left(y_i \middle| \frac{s_i + x_i}{2}, \frac{s_i - x_i}{2}\right) \\
=& p(y_i|s_i, x_i).
\end{aligned}$$

□

Owing to this parameterization, we expect that a univariate Skellam mean estimator of the form $\hat{x}_i = \mathbb{E}[x_i|y_i, s_i]$ will yield a satisfactory performance. The likelihood function of the observed filterbank coefficient y_i can now be rewritten as

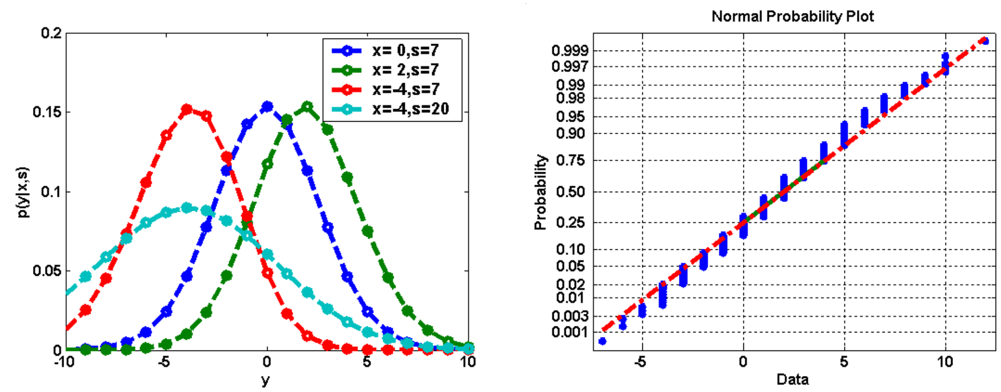
$$p(y_i|s_i, x_i) = e^{-s_i}\left(\frac{s_i + x_i}{s_i - x_i}\right)^{\frac{y_i}{2}} I_{y_i}\left(\sqrt{s_i^2 - x_i^2}\right).$$

Plots in Figure 1 illustrate $p(y_i|x_i, s_i)$. Note that it is in general asymmetric and heavy tailed—though as s increases, the Skellam distribution tends toward a Normal distribution. A random variable y_i with Skellam distribution has the following properties:

Proposition 3.2. *Let $p(y|x, s)$ be Skellam. Then:*

$$\begin{aligned}
\mathbb{E}[y|x, s] =& x^+ - x^- = x \\
\mathbb{E}[(y - x)^2|x, s] =& x^+ + x^- = s \\
\frac{d}{dx}p(y|x, s) =& \frac{p(y-1|x, s) - p(y+1|x, s)}{2} && (3) \\
\frac{d}{ds}p(y|x, s) =& \frac{p(y-1|x, s) + p(y+1|x, s)}{2} - p(y|x, s) && (4) \\
(y - x)p(y|x, s) =& s\frac{d}{dx}p(y|x, s) + x\frac{d}{ds}p(y|x, s). && (5)
\end{aligned}$$

The proofs for the mean and the variance are trivial; the proofs of (3-5) are provided in Appendix A.



(a) Probability Density Function $p(y|x,s)$. (b) Normal probability plot with $(x,s)=(2,7)$.

Figure 1. Illustrations of Skellam distribution

3.1. Unbiased Estimate of Risk and Shrinkage Operator

Let $\hat{x}_i$ indicate the Skellam mean estimator of x_i; we rewrite it as $\hat{x}_i(y_i) = y_i + \phi(y_i)$, where we hereafter omit the index i from $\hat{x}_i, x_i, y_i, s_i, t_i$ where understood. Its L^2 risk is written as:

$$\mathbb{E}\,\|\hat{x}(y) - x\|^2 = \mathbb{E}\,\|(y-x) + \phi(x)\|^2 = \mathbb{E}(y-x)^2 + 2\,\mathbb{E}(y-x)\phi(y) + \mathbb{E}\,\phi(y)^2. \tag{6}$$

Traditionally, first term is interpreted as a "constant" (i.e. independent of ϕ), and the last term is "computable" in terms of our observations. Using Hudson's results,[21] the second term can be rewritten as:

$$\begin{aligned}
\mathbb{E}(y^+ - x^+)\phi(y) &= \mathbb{E}\,y^+[\phi(y) - \phi(y-1)] \\
\mathbb{E}(y^- - x^-)\phi(y) &= \mathbb{E}\,y^-[\phi(y) - \phi(y+1)] \\
\mathbb{E}(y-x)\phi(y) &= \mathbb{E}(y^+ - x^+)\phi(y) - \mathbb{E}(y^- - x^-)\phi(y) \\
&= \mathbb{E}\,y^+[\phi(y) - \phi(y-1)] - \mathbb{E}\,y^-[\phi(y) - \phi(y+1)] \\
&= \mathbb{E}\,y\phi(y) - \mathbb{E}[y^+\phi(y-1) - y^-\phi(y+1)].
\end{aligned}$$

Thus the overall risk of the Skellam mean estimator is equivalent to

$$\mathbb{E}\,\|\hat{x}(y) - x\|^2 = \mathbb{E}(y-x)^2 + 2\,\mathbb{E}\,y\phi(y) - \mathbb{E}[y^+\phi(y-1) - y^-\phi(y+1)] + \mathbb{E}\,\phi(y)^2.$$

Note that with the exception of $(y-x)$, the risk of $\phi(\cdot)$ does not involve the latent variable $\boldsymbol{x}$. When $\hat{x}$ denotes a parametric estimator, its optimal parameter is said to be the minimizer of the following expression:

$$R(y) = 2\,\mathbb{E}\,y\phi(y) - \mathbb{E}[y^+\phi(y-1) - y^-\phi(y+1)] + \mathbb{E}\,\phi(y)^2.$$

Below, let $\hat{x}$ represent the standard soft thresholding operator, $\hat{x}(y) = \operatorname{sgn}(y)(|y| - \tau)_+$. Suppose we evaluate the risk in three separate scenarios:

- Suppose $|y| < \tau$—i.e. strictly below the threshold. Then $\phi(y) = -y$ and

$$\begin{aligned}
\mathbb{E}\left[\|\hat{x}(y) - x\|^2 \middle| y < \tau\right] &= \mathbb{E}(y-x)^2 + 2\,\mathbb{E}[-y^2 - (y^+(-y+1) - y^-(-y-1))] + \mathbb{E}\,y^2 \\
&= \mathbb{E}(y-x)^2 + 2\,\mathbb{E}[-y^2 - ((y^+ - y^-)(-y) + (y^+ + y^-))] + \mathbb{E}\,y^2 \\
&= \mathbb{E}(y-x)^2 - 2\,\mathbb{E}\,t + \mathbb{E}\,y^2
\end{aligned}$$

- Suppose $|y| > \tau$—i.e. strictly above the threshold. Then $\phi(y) = -\tau$ and

$$\begin{aligned}
\mathbb{E}\left[\|\hat{x}(y) - x\|^2 \middle| y > \tau\right] &= \mathbb{E}(y-x)^2 + 2\,\mathbb{E}[-\tau y - (-y^+\tau + y^-\tau)] + \mathbb{E}\,\tau^2 \\
&= \mathbb{E}(y-x)^2 + \mathbb{E}\,\tau^2
\end{aligned}$$

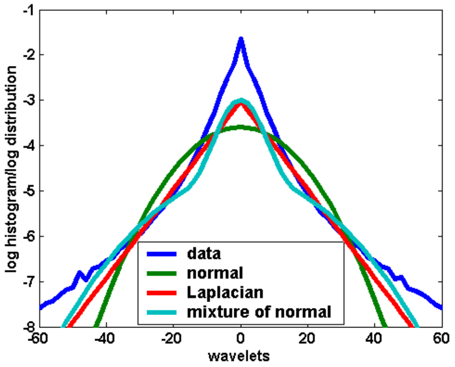


Figure 2. Log empirical histogram of the wavelet coefficients corresponding to an image signal, compared to candidate models for the distribution of coefficients.

- Suppose $|y| = \tau$—without loss of generality we assume $y > 0$. Then $\phi(y) = \phi(y+1) = -y = -\tau$, $\phi(y-1) = -y+1 = -\tau+1$, and

$$\begin{aligned}\mathbb{E}\left[\|\hat{x}(y)-x\|^2\Big|y=\tau\right] &= \mathbb{E}(y-x)^2 + 2\,\mathbb{E}[-\tau y - (y^+ + -y^+\tau + y^-\tau)] + \mathbb{E}\,\tau^2\\ &= \mathbb{E}(y-x)^2 - 2\,\mathbb{E}\,y^+ + \mathbb{E}\,\tau^2\end{aligned}$$

The overall estimated risk given a soft thresholding method is evaluated empirically using $\boldsymbol{y}$ *only* as:

$$\mathbb{E}\,\|\hat{x}(y_i)-x\|^2 \approx \sum (y_i-x_i)^2 + \sum_{|y_i|<\tau} y_i^2 + \sum_{|y_i|\geq\tau} \tau^2 - 2\sum_{|y_i|<\tau} t_i - 2\sum_{y_i=\tau} y_i^+ - 2\sum_{y_i=-\tau} y_i^-$$

In practice, $y_i = \pm\tau$ is an infrequent event—in fact, $y_i = \pm\tau$ is assumed to have zero mass in (1)—and thus we omit $y_i = \pm\tau$ to simplify the risk functional as:

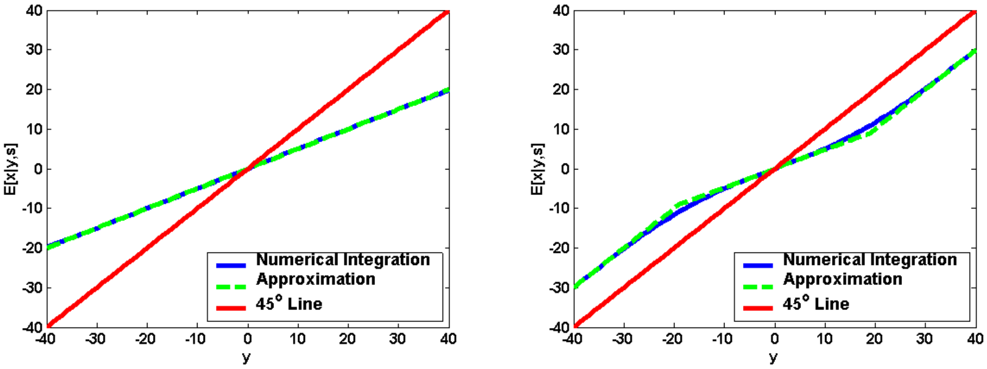
$$R_\tau(\boldsymbol{y}) = \sum_i \min(y_i^2, \tau^2) - 2\sum_{|y_i|<\tau} t_i, \tag{7}$$

and the optimal threshold is $\hat{\tau} = \arg\max_\tau R_\tau(\boldsymbol{y})$. An alternative derivation to the soft thresholding risk is provided in the appendix.

4. BAYESIAN APPROACH TO SKELLAM MEAN ESTIMATION

Suppose we adopt a Bayesian statistical framework—that is, we assume a prior distribution on the underlying transform coefficients, $\boldsymbol{x} = \boldsymbol{W}\boldsymbol{f}$. Determining the appropriate choice $p(\boldsymbol{x})$ is an active area of research in statistics and engineering that has shown promising results when applied to denoising for AWGN-type problems. For example, the canonical forms of heavy-tailed prior model distribution include the generalized Gaussian, Laplacian,[5] and finite and continuous mixtures of Gaussian.[1–3]—illustrated in Figure 2. As the goal of this paper is to estimate the transform coefficients $\boldsymbol{x}$ corresponding to the Poisson rate $\boldsymbol{f}$ given the "observed" coefficients $\boldsymbol{y} = \boldsymbol{W}\boldsymbol{f}$, we make no attempt to introduce additional insights into the prior model distribution below. Instead, we develop estimation techniques aimed at rigorous treatment of the underlying likelihood function, $p(\boldsymbol{y}|\boldsymbol{x})$.

Below, we derive the Bayesian form of the univariate Skellam mean estimator, where t_i is used in practice as



(a) Bayesian estimator with Normal prior.

(b) Bayesian estimator with Laplacian prior.

Figure 3. Bayesian estimator $\hat{x}(y,s) = \mathbb{E}[x|y,s]$ computed using (blue line) numerical integration and (green line) analytical approximation. The red 45^o line is shown for reference.

a plug-in estimate of s_i, which is unobservable:

$$\begin{aligned}\hat{x}_i(y_i,s_i) =& \mathbb{E}[x_i|y_i,s_i]\\ =& \int xp(x|y,s)dx\\ =& \int \frac{xp(y|x,s)p(x|s)}{p(y|s)}dx\\ =& \frac{\int xp(x|s)\left(\frac{s+x}{s-x}\right)^{\frac{y}{2}} I_y\left(\sqrt{s^2-x^2}\right)dx}{\int p(x|s)\left(\frac{s+x}{s-x}\right)^{\frac{y}{2}} I_y\left(\sqrt{s^2-x^2}\right)dx}\end{aligned} \tag{8}$$

To the best of the authors' knowledge, the existence of conjugate prior to Skellam distribution, which would make the above integrals analytically tractable, is not yet known. As a result, for a given choice of canonical prior model distributions $p(x)$ the integrals in $\hat{x}(\cdot,\cdot)$ must be evaluated *numerically.* The input-output "coring" functions corresponding to various choices of prior model distributions are shown in Figure 3.

However, we can show basic properties of the Bayesian estimator that admit meaningful interpretations and analytical approximations to $\hat{x}_i(y_i,s_i)$. We begin by exploring the likelihood function, $p(y_i|x_i,s_i)$, using (5):

$$xp(y|x,s) = yp(y|x,s) - s\frac{d}{dx}p(y|x,s) - x\frac{d}{ds}p(y|x,s) \tag{9}$$

$$\approx yp(y|x,s) - s\frac{d}{dx}p(y|x,s). \tag{10}$$

The approximation above is predicated on the "smoothness" of the likelihood function (i.e. $(1/2)[p(y-1|x,s)+p(y+1|x,s)] \approx p(y|x,s)$), which holds especially well in practice when s is large. This is useful in the context of the Bayesian estimation:

$$\mathbb{E}[x|y,s] = \frac{\int xp(y|x,s)dp(x|s)}{\int p(y|x,s)dp(x|s)} \approx y - s\frac{\int[\frac{d}{dx}p(y|x,s)]dp(x|s)}{p(y|s)} \tag{11}$$

Below, we consider two forms of canonical prior distribution: the Normal and the Laplacian.

Gaussian Prior Model

Supposing that $p(x|s) = p(x)$ is a truncated Normal distribution supported on $[-r,r]$, the following interpretation emerges.

Proposition 4.1. *Let $p(x|s)$ be a truncated Normal distribution on $[-r, r]$ and $p(y|x, s)$ be Skellam. Then the following approximation bound holds:*

$$\left| \frac{\int [\frac{d}{dx} p(y|x,s)] dp(x)}{p(y|s)} - \frac{\mathbb{E}[x|y,s]}{\sigma^2} \right| \le \left| \frac{ke^{-r^2}}{\sqrt{2\pi\sigma^2} p(y|s)} \right|.$$

where $k^{-1} = (2\pi\sigma^2)^{-\frac{1}{2}} \int_{-r}^{r} e^{-x^2/2\sigma^2} dx$. The proof of the proposition is provided in the appendix. Combining (11) and Proposition 4.1 we therefore rewrite the estimator:

$$\mathbb{E}[x|y,s] \approx \left(1 + \frac{s}{\sigma^2}\right)^{-1} y = \left(\frac{\sigma^2}{s + \sigma^2}\right) y. \tag{12}$$

Note that the above closely resembles a linear minimum mean squared error (LMMSE) estimator when s is interpreted as the variance of the noise.

Laplacian Prior Model

Supposing that $p(x|s) = p(x)$ is a truncated Laplacian distribution supported on $[-r, r]$, the following interpretation emerges.

Proposition 4.2. *Let $p(x|s)$ be truncated Laplacian on $[-r, r]$ and $p(y|x, s)$ be Skellam. Then the following approximation bound holds:*

$$\left| \frac{\int [\frac{d}{dx} p(y|x,s)] dp(x)}{p(y|s)} - \frac{p(x > 0|y,s) - p(x < 0|y,s)}{\sigma} \right| \le \left| \frac{ke^{-r}}{2\sigma p(y|s)} \right|.$$

The proof of the proposition is provided in the appendix. Combining (11) and Proposition 4.2 we therefore rewrite the estimator:

$$\mathbb{E}[x|y,s] \approx y + (s/\sigma)[p(x < 0|y,s) - p(x > 0|y,s)]. \tag{13}$$

A statistical interpretation of (13) can be illustrated via the posterior probability of x ($p(x|y,s)$) and Figure 4. Suppose $y > 0$. The area under this curve to the right of 0 is proportional to the amount of shrinkage towards zero, while the area to the left of 0 corresponds to the "counter-shrinkage" away from zero. It is therefore easy to see that large y corresponds to more shrinkage; large s contributes to more shrinkage because s/σ increases at a rate faster than $|p(x < 0|y,s) - p(x > 0|y,s)| < 1$. Thus the main features of the estimator are as follows:

- if y large, $p(x > 0|y) \approx 1$ and $\mathbb{E}[x|y,s] \approx y - (s/\sigma)$
- if y small, $p(x < 0|y) \approx 1$ and $\mathbb{E}[x|y,s] \approx y + (s/\sigma)$.
- the derivative of the estimator at $y = 0$ is not zero (verifiable via Figure 3).

Coincidentally, the second term of this approximate Bayesian estimator is similar to soft thresholding. Note, however, that the soft thresholding does not approximate the estimator well near the origin. In light of this, suppose we consider a computationally efficient piecewise linear approximation to (13):

$$\mathbb{E}[x|y,s] \approx \operatorname{sgn}(y) \max\Big(\lambda(s)|y|, |y| - s/\sigma\Big), \tag{14}$$

where $\lambda(s)$ is the derivative of the estimator at the origin ($y = 0$), which is easily computable. The following identities are readily verifiable from the definitions:

- $p(x > 0|y,s) + p(x < 0|y,s) = 1$,
- $\hat{x}(0,s) = 0$.

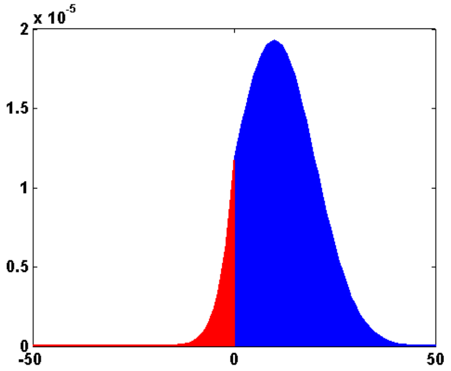


Figure 4. Posterior distribution of x, $p(x|y,s)$. When $y>0$, the area shaded in blue is proportional to the amount of shrinkage towards zero, while the area shaded in red corresponds to the "counter-shrinkage" away from zero.

- $\hat{x}(y,s) = -\hat{x}(-y,s)$.
- $\hat{x}(y,s) = y + (s/\sigma)[1-2p(x>0|y,s)]$

The slope of $\hat{x}$ evaluated at $y=0$ is then

$$\begin{aligned}\lambda(s) =& \frac{\hat{x}(1,s)-\hat{x}(-1,s)}{2} = \hat{x}(1,s)\\ =& 1 + (s/\sigma)[1-2p(x>0|y=1,s)]\end{aligned}$$

where the integral in $p(x>0|y=1,s)$ is independent of the observed data and therefore can be pre-computed offline (to be used as a small lookup table indexed by s).

Various analytical approximations to the exact Bayesian estimator $\mathbb{E}[x|y,s]$ of (8) are shown in Figure 3.

5. ACKNOWLEDGMENTS

The authors wish to express their gratitude to Drs. Alexander Berestov and C. C. Lee of Sony Electronics, Inc., for helping to make this work possible through numerous technical discussions and extensive programmatic support. This material is based upon work supported by Sony Electronics, Inc. and by the National Science Foundation under Grant DMS-0652743.

APPENDIX A. PROOF OF PROPOSITION 3.2

Skellam's original manuscript[11] gives us

$$\begin{aligned}\mathcal{F}p(y|x,s) =& \exp\left[-s + \left(\frac{s+x}{2}\right)e^{i\omega} + \left(\frac{s-x}{2}\right)e^{-i\omega}\right]\\ =& \exp\left[\left(\frac{s}{2}\right)(-2+e^{i\omega}+e^{-i\omega}) + \left(\frac{x}{2}\right)(e^{i\omega}-e^{-i\omega})\right] \qquad (15)\end{aligned}$$

where $\mathcal{F}$ denotes a Fourier transform with respect to y. Invoking linearity, we take the derivative in the Fourier domain:

$$\begin{aligned}\frac{d}{dx}p(y|x,s) =& \mathcal{F}^{-1}\frac{d}{dx}\mathcal{F}p(y|x,s)\\ =& \mathcal{F}^{-1}\frac{d}{dx}\exp\left[\left(\frac{s}{2}\right)(-2+e^{i\omega}+e^{-i\omega}) + \left(\frac{x}{2}\right)(e^{i\omega}-e^{-i\omega})\right]\\ =& \mathcal{F}^{-1}\left(\frac{e^{i\omega}}{2}-\frac{e^{-i\omega}}{2}\right)\exp\left[\left(\frac{s}{2}\right)(-2+e^{i\omega}+e^{-i\omega}) + \left(\frac{x}{2}\right)(e^{i\omega}-e^{-i\omega})\right]\\ =& \frac{p(y-1|x,s)-p(y+1|x,s)}{2},\end{aligned}$$

where the last step is a result of the phase shift property:

$$\mathcal{F}p(y \pm 1|x,s) = e^{\mp i\omega} \exp\left[\left(\frac{s}{2}\right)\left(-2 + e^{i\omega} + e^{-i\omega}\right) + \left(\frac{x}{2}\right)\left(e^{i\omega} - e^{-i\omega}\right)\right].$$

Similarly,

$$\begin{aligned}
\frac{d}{ds}p(y|x,s) =& \mathcal{F}^{-1}\frac{d}{ds}\mathcal{F}p(y|x,s) \\
=& \mathcal{F}^{-1}\frac{d}{ds}\exp\left[\left(\frac{s}{2}\right)\left(-2 + e^{i\omega} + e^{-i\omega}\right) + \left(\frac{x}{2}\right)\left(e^{i\omega} - e^{-i\omega}\right)\right] \\
=& \mathcal{F}^{-1}\left(\frac{e^{i\omega}}{2} + \frac{e^{-i\omega}}{2} - 1\right)\exp\left[\left(\frac{s}{2}\right)\left(-2 + e^{i\omega} + e^{-i\omega}\right) + \left(\frac{x}{2}\right)\left(e^{i\omega} - e^{-i\omega}\right)\right] \\
=& \frac{p(y-1|x,s) + p(y+1|x,s)}{2} - p(y|x,s).
\end{aligned}$$

Finally, basic identities of the modified Bessel function of the first kind yield the following:

$$\begin{aligned}
p(y|x,s) =& e^{-s}\left(\frac{s+x}{s-x}\right)^{\frac{y}{2}} I_y\left(\sqrt{s^2-x^2}\right) \\
=& e^{-s}\left(\frac{s+x}{s-x}\right)^{\frac{y}{2}}\left[I_{y-1}\left(\sqrt{s^2-x^2}\right) - I_{y+1}\left(\sqrt{s^2-x^2}\right)\right]\left(\frac{\sqrt{s^2-x^2}}{2y}\right) \\
=& \frac{s+x}{2y}p(y-1|x,s) - \frac{s-x}{2y}p(y+1|x,s) \\
=& \frac{s}{2y}\Big[p(y-1|x,s) - p(y+1|x,s)\Big] + \frac{x}{2y}\Big[p(y-1|x,s) + p(y+1|x,s)\Big].
\end{aligned} \tag{16}$$

Substituting (3) and (4), we complete the proof:

$$\begin{aligned}
yp(y|x,s) =& s\frac{d}{dx}p(y|x,s) + x\frac{d}{dx}p(y|x,s) + xp(y|x,s) \\
(y-x)p(y|x,s) =& s\frac{d}{dx}p(y|x,s) + x\frac{d}{dx}p(y|x,s).
\end{aligned}$$

APPENDIX B. PROOF OF PROPOSITIONS 4.1 AND 4.2

Proposition 4.1

Following the Stein's lemma strategy,[16–18] we take integration by parts:

$$\begin{aligned}
\frac{\int_{-r}^{r}\left(\frac{d}{dx}p(y|x,s)\right)dp(x)}{p(y|s)} =& \frac{p(x)p(y|x,s)\big|_{x=-r}^{r} - \int p(y|x,s)\left(\frac{d}{dx}p(x)\right)dx}{p(y|s)} \\
=& \frac{p(x=r)[p(y|x=r,s) - p(y|x=-r,s)] + (1/\sigma^2)\int xp(y|x,s)p(x)dx}{p(y|s)} \\
=& \frac{p(x=r)}{p(y|s)}[p(y|x=r,s) - p(y|x=-r,s)] + \frac{\mathbb{E}[x|y,s]}{\sigma^2},
\end{aligned}$$

where $(d/dx)p(x) = (-x/\sigma^2)p(x)$. Because $|p(y|x=r,s) - p(-y|x=r,s)| \leq 1$,

$$\left|\left[\frac{\int[\frac{d}{dx}p(y|x,s)]dp(x)}{p(y|s)}\right] - \left[\frac{\mathbb{E}[x|y,s]}{\sigma^2}\right]\right| \leq \left|\frac{ke^{-r^2}}{\sqrt{2\pi\sigma^2}p(y|s)}\right|.$$

Proposition 4.2

Recall (15) and $p(x) = (k/2\sigma)e^{-|x|/\sigma}$. Then the quantity,

$$
\begin{aligned}
&\frac{-\operatorname{sgn}(x)p(y|x,s)}{\sigma} + \frac{d}{dx}p(y|x,s) \\
&\quad = \frac{-\operatorname{sgn}(x)p(y|x,s)}{\sigma} + \frac{p(y-1|x,s) - p(y+1|x,s)}{2} \\
&\quad = \mathcal{F}^{-1}\left\{\left(\frac{-\operatorname{sgn}(x)}{\sigma} + \frac{e^{i\omega}}{2} - \frac{e^{-i\omega}}{2}\right)\exp\left[\left(\frac{s}{2}\right)(-2 + e^{i\omega} + e^{-i\omega}) + \left(\frac{x}{2}\right)(e^{i\omega} - e^{-i\omega})\right]\right\}
\end{aligned}
$$

is integrable in the following sense:

$$
\begin{aligned}
&\mathcal{F}^{-1}\int_{-r}^{r}\mathcal{F}\left\{\frac{-\operatorname{sgn}(x)p(y|x,s)}{\sigma} + \frac{p(y-1|x,s)}{2} - \frac{p(y+1|x,s)}{2}\right\}\left(\frac{k}{2\sigma}\right)e^{-\frac{|x|}{\sigma}}dx \\
&\quad = \mathcal{F}^{-1}\int_{0}^{r} + \int_{-r}^{0}\left(\frac{k}{2\sigma}\right)\left(\frac{-\operatorname{sgn}(x)}{\sigma} + \frac{e^{i\omega}}{2} - \frac{e^{-i\omega}}{2}\right)\exp\left[\left(\frac{s}{2}\right)(-2 + e^{i\omega} + e^{-i\omega}) + \left(\frac{x}{2}\right)(e^{i\omega} - e^{-i\omega}) - \frac{|x|}{\sigma}\right]dx \\
&\quad = \mathcal{F}^{-1}\left(\frac{k}{2\sigma}\right)\exp\left[\left(\frac{s}{2}\right)(-2 + e^{i\omega} + e^{-i\omega}) + \left(\frac{x}{2}\right)(e^{i\omega} - e^{-i\omega}) - \frac{|x|}{\sigma}\right]\Bigg|_{x=-r}^{r} \\
&\quad = \left(\frac{ke^{-r}}{2\sigma}\right)\left[e^{-s}\left(\frac{s+r}{s-r}\right)^{\frac{y}{2}} I_y(\sqrt{s^2 - r^2}) - e^{-s}\left(\frac{s-r}{s+r}\right)^{\frac{y}{2}} I_y(\sqrt{s^2 - r^2})\right] \\
&\quad = \left(\frac{ke^{-r}}{2\sigma}\right)[p(y|x=r,s) - p(-y|x=r,s)].
\end{aligned}
$$

Because $|p(y|x=r,s) - p(-y|x=r,s)| \leq 1$,

$$
\left|\left[\frac{\int[\frac{d}{dx}p(y|x,s)]dp(x)}{p(y|s)}\right] - \left[\frac{\int \operatorname{sgn}(x)p(y|x,s)dp(x)}{\sigma p(y|s)}\right]\right| \leq \left|\frac{ke^{-r}}{2\sigma p(y|s)}\right|.
$$

The above quantity decays very rapidly as r gets large, and for all practical purposes, this can be approximated as zero. Expanding the integral completes the proof:

$$
\begin{aligned}
\frac{\int \operatorname{sgn}(x)p(y|x,s)dp(x)}{\sigma p(y|s)} &= \int_{0}^{r} - \int_{-r}^{0}\frac{p(y|x,s)p(x)}{p(y|s)}dx \\
&= \int_{0}^{r} - \int_{-r}^{0}p(x|y,s)dx \\
&= p(x>0|y,s) - p(x<0|y,s).
\end{aligned}
$$

REFERENCES

1. K. Hirakawa and X.-L. Meng, “An empirical Bayes EM-wavelet unification for simultaneous denoising, interpolation, and/or demosaicing,” *IEEE Int’l Conf Image Processing* , October 2006.
2. J. Portilla, V. Strela, M. Wainwright, and E. Simoncelli, “Image denoising using scale mixtures of Gaussians in the wavelet domain,” *Image Processing, IEEE Transactions on* **12**(11), pp. 1338–1351, 2003.
3. M. S. Crouse, R. D. Nowak, and R. G. Baraniuk, “Wavelet-based statistical signal processing using hidden Markov models,” *IEEE Int’l Conf on Acoustics, Speech, and Signal Processing* , 1998.
4. A. Pizurica, W. Philips, I. Lemahieu, and M. Acheroy, “A joint inter- and intrascale statistical model for Bayesian wavelet based image denoising,” *IEEE Trans. Image Processing* **11**(5), pp. 545–557, 2002.
5. L. Sendur and I. W. Selesnick, “Bivariate shrinkage functions for wavelet-based denoising exploiting interscale dependency,” *IEEE Trans. Signal Processing* **50**, 2002.
6. A. Papoulis and S. U. Pillai, *Probability, Random Variables, and Stochastic Processes*, McGraw Hill, 4th ed., 2002.

7. A. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-time signal processing*, Prentice-Hall, Inc., 1989.
8. P. Fryzlewicz and G. P. Nason, "A Haar-Fisz algorithm for Poisson intensity estimation," *J. Comp. Graph. Stat.* **13**, pp. 621–638, 2004.
9. M. Fisz, "The limiting distribution of a function of two independent random variables and its statistical application," *Colloquium Mathematicum* **3**, pp. 138–146, 1955.
10. C. Kervrann and A. Trubuil, "An adaptive window approach for poisson noise reduction and structure preserving in confocal microscopy," in *Biomedical Imaging: Macro to Nano, 2004. IEEE Int'l Symposium on*, **1**, pp. 788–791, 2004.
11. J. G. Skellam, "The frequency distribution of the difference between two Poisson variates belonging to different populations," *Journal of the Royal Statistical Society* **109**(3), p. 296, 1946.
12. D. Karlis and I. Ntzoufras, "Analysis of sports data using bivariate poisson models," *Journal of the Royal Statistical Society: Series D* **52**(3), pp. 381—393, 2003.
13. D. Karlis and I. Ntzoufras, "Bayesian analysis of the differences of count data," *Statistics in Medicine* (25), pp. 1885—1905, 2006.
14. Y. Hwang, J. Kim, and I. Kweon, "Sensor noise modeling using the Skellam distribution: Application to the color edge detection," *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on* , pp. 1–8, 2007.
15. Y. Hwang, I. Kweon, and J. Kim, "Color edge detection using the Skellam distribution as a sensor noise model," *SICE, 2007. Annual Conference* , pp. 1972–1979, 2007.
16. C. Stein, "Confidence sets for the mean of a multivariate normal distribution (with discussion)," *J. Roy. Statist. Soc. Ser. B* **24**, pp. 265–296, 1962.
17. C. Stein, "Inadmissibility of the usual estimator for the mean of a multivariate normal distribution," in *Proc. Third Berlzeley Symp. Math. Statist. Probab.*, **1**, pp. 197–206, 1956.
18. C. Stein, "Estimation of the mean of a multivariate normal distribution," *The Annals of Statistics* **9**, pp. 1135–1151, 1981.
19. D. L. Donoho and I. M. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage," *Journal of the American Statistical Association* **90**(432), pp. 1200–1224, 1995.
20. I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, Academic Press, New York, 5th ed., 1994.
21. H. M. Hudson, "Adaptive estimators for simultaneous estimation of poisson means," *The Annals of Statistics* **13**, pp. 246–261, March 1985.