

RELAXING THE GAUSSIAN ASSUMPTION IN SHRINKAGE AND SURE IN HIGH DIMENSION

BY MAX FATHI^{1,a}, LARRY GOLDSTEIN^{2,b}, GESINE REINERT^{3,c} AND
ADRIEN SAUMARD^{4,d}

¹Université Paris Cité and Sorbonne Université, CNRS, Laboratoire Jacques-Louis Lions & Laboratoire de Probabilités, Statistique et Modélisation, ^amfathi@lpsm.paris

²Department of Mathematics, University of Southern California, ^blarry@math.usc.edu

³Department of Statistics, University of Oxford, ^creinert@stats.ox.ac.uk

⁴Université de Rennes, Ensai, CREST-UMR 9194, ^dadrien.saumard@ensai.fr

Shrinkage estimation is a fundamental tool of modern statistics, pioneered by Charles Stein upon his discovery of the famous paradox involving the multivariate Gaussian. A large portion of the subsequent literature only considers the efficiency of shrinkage, and that of an associated procedure known as Stein’s Unbiased Risk Estimate, or SURE, in the Gaussian setting of that original work. We investigate what extensions to the domain of validity of shrinkage and SURE can be made away from the Gaussian through the use of tools developed in the probabilistic area now known as Stein’s method. We show that shrinkage is efficient away from the Gaussian under very mild conditions on the distribution of the noise. SURE is also proved to be adaptive under similar assumptions, and in particular in a way that retains the classical asymptotics of Pinsker’s theorem. Notably, shrinkage and SURE are shown to be efficient under mild distributional assumptions, and particularly for general isotropic log-concave measures.

1. Introduction. The breakthrough, counterintuitive results of the works [41] and [26] showed that the “natural” estimate of the unknown mean $\theta \in \mathbb{R}^d$ of an observation X having the normal distribution $\mathcal{N}_d(\theta, \sigma^2 \text{Id})$ in dimensions $d \geq 3$ is not admissible under mean squared error loss. In particular, with $\|\cdot\|$ denoting the Euclidean norm, for $d \geq 3$ it was demonstrated that for

$$(1) \quad S_\lambda(X) = X \left(1 - \frac{\lambda}{\|X\|^2} \right) \quad \text{for } \lambda \geq 0$$

there exists a range of positive values for λ for which $S_\lambda(X)$ has a strictly smaller mean squared error than $S_0(X)$. Here, by the properties of the Gaussian, X could represent the mean of an independent sample from this same Gaussian distribution with σ^2 properly rescaled.

In [42], related ideas, and in particular the use of Stein’s lemma, were applied to construct what is now known as SURE, for Stein’s Unbiased Risk Estimate, that provides an unbiased estimator for the mean squared error of a nearly arbitrary estimator of a multivariate mean, again in the Gaussian context. For $f : \mathbb{R}^d \rightarrow \mathbb{R}^d$, let ∇f and $\nabla \cdot f$ denote the Jacobian matrix, and divergence of f , respectively; precisely, with ∂_j denoting taking the partial derivative with respect to the j th coordinate variable, $[\nabla f]_{i,j} = \partial_j f_i$. With ν generally denoting the distribution of X , which for now is the normal $\mathcal{N}_d(\theta, \sigma^2 \text{Id})$, let $W^{1,2}(\nu)$ denote the natural (weighted) Sobolev space induced by the (squared) Sobolev norm

$$(2) \quad \|f\|_{W^{1,2}(\nu)}^2 := \|f\|_{L^2(\nu)}^2 + \|\nabla f\|_{L^2(\nu)}^2,$$

Received April 2020; revised March 2022.

MSC2020 subject classifications. 62F12, 62F35.

Key words and phrases. Shrinkage estimation, Stein kernel, zero bias, unbiased risk estimation.

where the second term is the usual (squared) Hilbert–Schmidt norm induced by the scalar product $\langle A, B \rangle = \text{Tr}(AB^\top)$ on matrices A and B (see, for instance, [13]). Stein’s identity gives the characterization of the multivariate normal distribution that $X \sim \mathcal{N}_d(\theta, \Sigma)$ if and only if

$$(3) \quad E[\langle X - \theta, f(X) \rangle] = E[\langle \Sigma, \nabla f(X) \rangle] \quad \text{for all } f \in W^{1,2}(\nu).$$

In particular, via (3), for $f \in W^{1,2}(\nu)$ and $X \sim \mathcal{N}_d(\theta, \sigma^2 \text{Id})$ we have that

$$(4) \quad \begin{aligned} \text{SURE}(f, X) &:= d\sigma^2 + \|f(X)\|^2 + 2\sigma^2 \nabla \cdot f(X) \\ &\text{is unbiased for the risk of } S(X) = X + f(X), \end{aligned}$$

that is, unbiased for the expectation of $\|S(X) - \theta\|^2$. In particular, taking

$$(5) \quad f(x) = -\lambda g_0(x) \quad \text{where } g_0(x) = \frac{x}{\|x\|^2}$$

in (4) gives an unbiased estimator for the risk of the shrinkage estimator (1).

Since shrinkage estimation and SURE first appeared, they have been extensively studied in the statistical literature and applied in practice in many contexts; see, for instance [4, 7, 16, 46]. Regarding shrinkage, previous result for non-Gaussian distributions have appeared in the works [22] and [9, 12, 40] that consider the estimation of high-dimensional covariance matrices under spherically and elliptically symmetric distributional assumptions. Compared to those works, an advantage of our approach is that a number of our main results completely avoid any assumption of symmetry.

In addition, the work [19] considers shrinkage estimation of the mean θ based on the observation $X = Y + \theta$ under the assumption that $d \geq 3$, $E[Y] = 0$, $E[\|Y\|^2] < \infty$ and that there exists a (possibly randomized) stopping time t for an \mathbb{R}^d valued Brownian motion $B_{s \geq 0}$ such that the distribution of B_t is that of Y . However, the proof of the main result of [19] appears to be in error, in that it does not take into account that the stopping time involved depends on the path, and that its variation must be taken into account when taking a derivative with respect to the parameter ϵ that controls the magnitude of the drift of the perturbed Brownian motion constructed when deriving [19], equation (3).

Regarding the use of SURE in non-Gaussian settings, [17] extends SURE to exponential families by exploiting the fact that in the natural parametrization the score function is linear in the unknown θ , allowing linear functions of this unknown to be unbiasedly estimated using quantities that do not depend on it. The approach taken in [17] is unrelated to the methods we consider, and the results obtained are presently not subsumed by ours.

Assuming independence of the coordinates, [32] considers consistency of SURE and of Stein shrinkage type estimators in the context of linear estimation, and makes an appealing link with generalized cross-validation. Precise comparison with our results are presented in Remark 6.3. This present work makes the case that SURE can be extended beyond the currently known settings. Though unbiased for the Gaussian, SURE can be applied in many cases “as is” at the cost of a bias of order small enough to be able to, say, consistently choose good tuning parameters. In particular, we show that under our conditions SURE remains adaptive, in that the classical asymptotics of Pinsker’s theorem for the Gaussian case still apply. We propose to distinguish this estimate by using the term ASSURE for the non-Gaussian cases where the procedure is Approximately the Same as SURE.

We verify properties of our proposed extensions using tools having their origins in Stein’s method, in particular, Stein kernels and the zero bias distribution. We present a review of shrinkage and SURE in the Gaussian case, followed by background needed for the application of the methods we apply; technical results used for the zero bias technique in multi-dimension

are compiled in Section 5. In Section 3, we present a number of non-Gaussian models, to our knowledge not previously discussed in the literature, under which shrinkage is shown to be advantageous. In Section 4, we assess SURE in non-Gaussian cases, concluding that the bias incurred by applying the estimate used in the Gaussian case can be small enough as to make this estimate useful in many instances. In addition, in Section 4.2 we consider the use of SURE for soft-thresholding, and in Section 4.3 consider the adaptivity of the shrinkage estimator, and extensions of Pinsker's theorem, for non-Gaussian cases. Section 6 gives some technical results needed in Sections 3 and 4 on the boundedness in mean of inverse norms.

The main proofs of our results and some technical details on a few illustrative examples appear in the Supplementary Material [21]. In addition, in part F of the Supplement we give conditions under which our Stein kernel methods may be applied to functions other than the special case of $g_0(\mathbf{x})$ in (5) that is specific to shrinkage, and which is handled for that case by Assumption 3.1.

Though our results cover classes of distributions previously not treated in the literature, such nonelliptical or discrete distributions, some specific applications of our methods give an introduction to our results. For instance, Examples 3.2 and 3.4 cover the multivariate Student t distribution, Example 3.3 the uniform distribution on the sphere S^{d-1} and with support over an ellipsoid, and Example 3.5 considers corrupted Gaussian observations.

For instance, for \mathbf{X} following a d -dimensional multivariate Student t distribution with $k \geq 5$ degrees of freedom, having unknown mean $\boldsymbol{\theta}$ and known covariance matrix Σ with largest eigenvalue κ , for even dimensions $d = 2m \geq 6$, we apply two approaches, yielding the two different bounds

$$24\lambda \sqrt{\frac{2d^2(k+2)}{(d-2)(d-4)(d+k-2)k(k-4)}} \quad \text{and} \quad 16\lambda \frac{(d+k-2)}{(d-2)k}$$

on the *excess risk*, that is, the mean squared error above its value in the Gaussian case. The first bound, from Example 3.2, uses a Stein kernel and Theorem 3.2, while the second, from Example 3.4, applies a zero bias approach in conjunction with Theorem 3.4. When $\lambda \in [0, 2(\text{Tr}(\Sigma) - 2\kappa)]$ the risk of S_λ is no larger than that of S_0 asymptotically in the first case when $kd \rightarrow \infty$, and for the second when $k \rightarrow \infty$ as $d \rightarrow \infty$. As in general it may be the case that only one of the two results applied here may be invoked, having access to both these approaches is advantageous even though the conclusions drawn for this particular case are nearly the same.

In the following, densities of random vectors are with respect to Lebesgue measure, and when \mathbf{X} has measure ν we will refer to the measure of $\mathbf{X} + \boldsymbol{\theta}$ as the translation of ν by $\boldsymbol{\theta}$.

2. Stein's identity and two extensions. Stein's identity [42], also known as Stein's lemma, for characterizing the one-dimensional normal distribution states that a random variable X has law $\mathcal{N}(\theta, \sigma^2)$ if and only if

$$(6) \quad E[(X - \theta)f(X)] = \sigma^2 E[f'(X)] \quad \text{for all } f \in \mathcal{F},$$

where \mathcal{F} is the class of all real valued functions that are absolutely continuous on compact intervals, and for which the expectation on the left-hand side of (6) exists; an extension to d dimensions is given in (3). We consider two generalizations of Stein's lemma that will be used for the relaxation of the normal assumptions in Shrinkage and SURE.

One way that Stein's lemma may be generalized in one dimension for a mean zero random variable X is through the use of a Stein kernel T , a random variable for which

$$E[Xf(X)] = E[Tf'(X)] \quad \text{for all } f \in \mathcal{F}.$$

Stein kernels were first introduced in [6], and further developed in the univariate setting in [10]. By replacing T by $E[T|X]$, we may assume that T is some function of X .

When X has mean θ and $T_{X-\theta}$ is the Stein kernel of $X - \theta$, we obtain

$$\begin{aligned} E[(X - \theta)f(X)] &= E[(X - \theta)f((X - \theta) + \theta)] \\ &= E[T_{X-\theta}f'((X - \theta) + \theta)] = E[T_{X-\theta}f'(X)]. \end{aligned}$$

Hence, if we are given $X = Y + \theta$ for an unknown θ and some mean zero random variable Y with known distribution, we usually cannot compute the Stein kernel $T_{X-\theta}$ for $X - \theta$ without knowledge of θ . However, under natural assumptions we can get estimates on norms of $T_{X-\theta}$ that are uniform in θ , for example, as in the setting of [13].

Another way to generalize Stein’s lemma to non-Gaussian cases, following [24] and [15], is to use the fact that for any random variable X with finite, nonzero variance σ^2 and mean θ ; the X -zero bias distribution X^* exists, which is characterized by the condition that

(7)
$$E[(X - \theta)f(X)] = \sigma^2 E[f'(X^*)] \quad \text{for all } f \in \mathcal{F}.$$

Hence, Stein’s lemma (6) can be restated as saying that the univariate normal distributions are the unique fixed points of the zero bias transformation that produces the distribution of X^* from that of X .

We highlight a relation between the zero bias distributions in the centered and noncentered cases by noting that if we define X^* via (7) restricting to the case where $\theta = 0$ then for the general case we obtain (7) by letting

(8)
$$X^* := (X - \theta)^* + \theta.$$

This distinction is important. With $=_d$ denoting equality in distribution, if we are given that X is from a location family, specifically, that if $X =_d Y + \theta$ where the distribution of some mean zero variable Y is specified but θ is not, then similar to this phenomenon for Stein kernels, though we may sample from Y^* , we are not able to sample from X^* without knowledge of θ .

The intricacy of the Stein kernel T and the zero bias distribution is not illustrated in the normal case, as there T is simply the variance, and the transformed distribution unchanged from the original, respectively. We now move on to multivariate generalizations of Stein kernels and the zero bias distribution.

2.1. Multidimensional Stein kernels. Stein kernels can be defined in the multivariate setting of vectors with dependent coordinates. This notion, originating in [10], is at the core of the Nourdin–Peccati approach to Stein’s method [35], making a powerful link to Malliavin calculus. Given a random vector $X \in \mathbb{R}^d$ with mean θ and distribution ν , which is absolutely continuous with respect to Lebesgue measure on \mathbb{R}^d , a Stein kernel $T_{X-\theta}$ for the mean-zero vector $X - \theta$ is a matrix-valued function such that

(9)
$$E[(X - \theta, f(X))] = E[(T_{X-\theta}, \nabla f(X))] \quad \text{for all } f \in W^{1,2}(\nu).$$

We remark that other works on Stein kernels may use other classes of test functions; see, for example, the discussion in [33]. In addition, we only consider situations where the Stein kernel has a finite second moment, so that the right-hand side is always finite for test functions in $W^{1,2}(\nu)$.

As in the univariate setting, the Stein characterization (3) of the normal distribution translates as saying that a random vector X has a Gaussian distribution with mean θ and covariance matrix Σ iff $X - \theta$ admits Σ as a Stein kernel.

Construction of multidimensional Stein kernels as in (9) has been considered, for example, in [13, 20, 33]. Existence and uniqueness of Stein kernels in higher dimensions are not in general guaranteed. In [13], it is shown that if ν is centered and satisfies a Poincaré inequality, then there is a Stein kernel that is the gradient of an element of $W^{1,2}_0(\nu)$, the set of functions

in $W^{1,2}(\nu)$ with ν -mean zero, and it is unique in that class. When the components of $\mathbf{Y} = (Y_1, \dots, Y_d)$ are independent, with mean zero, finite variances and admit Stein kernels T_i , $i = 1, \dots, d$, [33], Example 3.9, shows that the diagonal matrix $T = \text{diag}(T_1, \dots, T_d)$, satisfies (9) with $\boldsymbol{\theta} = 0$.

REMARK 2.1. If $T_{\mathbf{Y}}$ is a Stein kernel for a centered isotropic random vector \mathbf{Y} and A is an invertible matrix, then

$$(10) \quad E[\langle A\mathbf{Y}, \mathbf{f}(A\mathbf{Y}) \rangle] = E[\langle AT_{\mathbf{Y}}A^{\top}, \nabla \mathbf{f}(A\mathbf{Y}) \rangle]$$

so that $\mathbf{y} \longrightarrow AT_{\mathbf{Y}}(A^{-1}\mathbf{y})A^{\top}$ is a Stein kernel for $A\mathbf{Y}$. In particular, this transformation with $A = \text{Cov}(\mathbf{Y})^{-1/2}$ allows us to reduce many statements for nonisotropic random vectors \mathbf{Y} to the isotropic case, as long as the covariance matrix of \mathbf{Y} is invertible.

EXAMPLE 2.1. We say that an absolutely continuous \mathbb{R}^d valued random vector \mathbf{X} has a multivariate elliptical distribution $E_d(\boldsymbol{\theta}, \Upsilon, \phi)$ if it admits a density of the form

$$p(\mathbf{x}) = \kappa |\Upsilon|^{-1/2} \phi\left(\frac{1}{2}(\mathbf{x} - \boldsymbol{\theta})^{\top} \Upsilon^{-1}(\mathbf{x} - \boldsymbol{\theta})\right), \quad \mathbf{x} \in \mathbb{R}^d,$$

for $\phi : [0, \infty) \rightarrow [0, \infty)$ a measurable function called a *density generator*, $\boldsymbol{\theta} \in \mathbb{R}^d$ the location parameter, κ the normalizing constant and Υ a symmetric positive definite $d \times d$ dispersion matrix. Here, we assume that the model is chosen such that $\boldsymbol{\theta}$ is the mean vector and Υ is the covariance matrix Σ of \mathbf{X} ; see, for example, [29] for suitable conditions.

The cases $E_d(0, \text{Id}, \phi)$ are the *spherical distributions*. They are centred and isotropic, and hence (10) applies, so that if T_{ϕ} is a Stein kernel for $E_d(0, \text{Id}, \phi)$ then $T_{\mathbf{X}-\boldsymbol{\theta}}(\mathbf{x}) = \Sigma^{1/2} T_{\phi}(\Sigma^{-1/2}(\mathbf{x} - \boldsymbol{\theta})) \Sigma^{1/2}$ is a Stein kernel for $E_d(\boldsymbol{\theta}, \Sigma, \phi)$; see also [33]. In particular, by [33] and Proposition 2 in [29], a Stein kernel for $E_d(\boldsymbol{\theta}, \Sigma, \phi)$ is given by

$$(11) \quad T_{\mathbf{X}-\boldsymbol{\theta}}(\mathbf{x}) = \left(\frac{1}{\phi((\mathbf{x} - \boldsymbol{\theta})^{\top} \Sigma^{-1}(\mathbf{x} - \boldsymbol{\theta})/2)} \int_{(\mathbf{x}-\boldsymbol{\theta})^{\top} \Sigma^{-1}(\mathbf{x}-\boldsymbol{\theta})/2}^{+\infty} \phi(u) du \right) \Sigma.$$

Moving forward, when considering the shrinkage estimator (1) for non-Gaussian models using Stein kernels, for integrability we will require, unless other conditions are explicitly mentioned, that the following assumption is in force.

ASSUMPTION 2.1. The function $\mathbf{g}_0(\mathbf{x}) = \mathbf{x}/\|\mathbf{x}\|^2$ is an element of $W^{1,2}(\nu)$.

We note for later use that

$$(12) \quad \nabla \mathbf{g}_0(\mathbf{x}) = \frac{1}{\|\mathbf{x}\|^2} \text{Id} - \frac{2}{\|\mathbf{x}\|^4} \mathbf{x} \mathbf{x}^{\top}.$$

Lemma 2.1 provides the following simple sufficient condition for the satisfaction of Assumption 2.1. Its proof is given in Supplement A.

LEMMA 2.1. When $d \geq 5$, Assumption 2.1 is satisfied by the measure ν when it has a density bounded almost everywhere in some neighborhood of the origin, and by the translates of ν by any $\boldsymbol{\theta} \in \mathbb{R}^d$ when ν has a density bounded almost everywhere from above.

2.2. Multidimensional zero-bias transform. Let ν be a given probability measure on \mathbb{R}^d possessing second moments, and for $1 \leq i, j \leq d$ let $\sigma_{ij} = \text{Cov}(Y_i, Y_j)$ when $\mathcal{L}(Y) = \nu$. For i, j such that $\sigma_{ij} \neq 0$, consider probability measures ν^{ij} on \mathbb{R}^d depending on ν , and using notation parallel to (2), define the Sobolev-like norm and its corresponding function space, respectively, by

$$\|f\|_{W_z^{1,2}(\nu)}^2 := \|f\|_{L^2(\nu)}^2 + \sum_{i,j:\sigma_{ij}\neq 0}^d \|\partial_j f_i\|_{L^2(\nu^{ij})}^2 \quad \text{and} \quad W_z^{1,2}(\nu) = \{f : \|f\|_{W_z^{1,2}(\nu)}^2 < \infty\}.$$

The multivariate extension for zero biasing below in (13) extends and complements the generalization in [25]. For our extension, given a mean zero random vector Y in \mathbb{R}^d with positive definite covariance matrix Σ having entries $\sigma_{ij} = \text{Cov}(Y_i, Y_j)$, we say the collection of vectors $\{Y^{ij} : i, j \text{ such that } \sigma_{ij} \neq 0\}$ in \mathbb{R}^d has the multivariate Y -zero bias distribution when

$$(13) \quad E[\langle Y, f(Y) \rangle] = E\left[\sum_{i,j=1}^d \sigma_{ij} \partial_j f_i(Y^{ij})\right] =: E[\langle \Sigma, \nabla f(Y^*) \rangle] \quad \text{for all } f \in W_z^{1,2}(\nu),$$

where in the second equality we define $\nabla f(Y^*)$ to be the matrix with i, j th entry $\partial_j f_i(Y^{ij})$. When this identity is satisfied, we say that the zero bias vectors of Y exist.

Though in point 5 of Proposition 5.1, and in Example 3.3 we consider the zero bias approach where (13) holds for a distribution with nondiagonal covariance matrix below, in view of Remark 2.1, we focus primarily on the case where $\Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_d^2)$, an invertible diagonal matrix. In this instance, the collection of zero bias vectors appearing in the identity (13) reduce to the d vectors $Y^i = Y^{ii}$, for which, now also letting $\sigma_i^2 = \sigma_{ii}$, satisfy

$$(14) \quad E[\langle Y, f(Y) \rangle] = E\left[\sum_{i=1}^d \sigma_i^2 \partial_i f_i(Y^i)\right] \quad \text{for all } f \in W_z^{1,2}(\nu).$$

Part 1 of Proposition 5.1 in Section 5 shows that the zero bias vectors exist for any mean zero Y with nonsingular diagonal covariance matrix if and only if

$$(15) \quad E[Y_i | Y_j, j \neq i] = 0 \quad \forall i = 1, \dots, d,$$

and, under this condition, provides a construction. Part 2 of Proposition 5.1 specifies the support of the zero bias vectors, Part 3 considers independent sums, Part 4 handles mixtures, Part 5 provides the existence of the zero bias vectors for a class of distributions with non-diagonal covariance matrices and Part 6 considers the special case where Y possesses a density function.

Generally, to encompass vectors X with arbitrary means θ , extending (7) and (8), (13) implies that

$$(16) \quad E[\langle X - \theta, f(X) \rangle] = E\left[\sum_{i,j=1}^d \sigma_{ij} \partial_j f_i(X^{ij})\right] \quad \text{for } X^{ij} = (X - \theta)^{ij} + \theta.$$

EXAMPLE 2.2. Following on from Example 2.1, for $Y \sim E_d(0, \text{Id}, \phi)$ it is shown implicitly in Proposition 2 of [29] that the collection $Y^i = Y^* \sim E_d(0, \text{Id}, \Phi)$, for $i = 1, \dots, d$ has the Y -zero bias distribution, where $\Phi(x) = \int_x^\infty \phi(u) du$. Example 3.3, which considers the spherical distribution resulting by placing uniform measure on the surface of a sphere, is not covered by the referenced results as it is not absolutely continuous with respect to Lebesgue measure on \mathbb{R}^d .

When $\sigma_{ij} \geq 0$ for all $1 \leq i, j \leq d$, the right-hand side of (13) may be written more compactly as a mixture via the use of a pair of random indices (I, J) , independent of $\{Y^{ij}, i, j = 1, \dots, d\}$, with distribution

$$P(I = i, J = j) = \frac{\sigma_{ij}}{\sigma^2} \quad \text{where } \sigma^2 = \text{Var}\left(\sum_{i=1}^d Y_i\right) = \sum_{i,j=1}^d \sigma_{ij}.$$

Then, starting with the first equality of (13), we obtain

$$\begin{aligned} E[\langle Y, f(Y) \rangle] &= E\left[\sum_{i,j=1}^d \sigma_{ij} \partial_j f_i(Y^{ij})\right] \\ &= \sigma^2 E\left[\sum_{i,j=1}^d P(I = i, J = j) \partial_j f_i(Y^{ij})\right] = \sigma^2 E[\partial_J f_I(Y^{IJ})]. \end{aligned}$$

In particular, taking $g(y) = \sum_i y_i$, the sum of the coordinates of $y \in \mathbb{R}^d$, $W = g(Y)$ and $f(y) = (f(g(y)), \dots, f(g(y)))$ for smooth f yields

$$(17) \quad E[Wf(W)] = E[\langle Y, f(Y) \rangle] = \sigma^2 E[f'(W^{IJ})],$$

demonstrating that W^{IJ} , the sum of the coordinates of Y^{IJ} , has the W -zero biased distribution. We note that the condition that σ_{ij} be nonnegative always holds when Y has a diagonal covariance matrix.

As was done for Stein kernels by Assumption 2.1, here we shall adopt Assumption 2.2 to guarantee that the zero bias Stein identity can be applied to the function $g_0(x)$ that is used in the shrinkage estimator.

ASSUMPTION 2.2. The function $g_0(x) = x/\|x\|^2$ is an element of $W_z^{1,2}(\nu)$.

Similar to the sufficient condition provided by Lemma 2.1 for the satisfaction of Assumption 2.1 when applying kernels, here we provide some simple conditions that guarantee the validity of Assumption 2.2; we restrict to the diagonal covariance case. The proof of Lemma 2.2 can be found in Supplement A.

LEMMA 2.2. Let $d \geq 5$ and ν the measure of a mean zero distribution with finite second moment, that satisfies (15). Then Assumption 2.2 is satisfied for the measure ν and all its translates when ν has a density $p(y)$ such that for each $i = 1, \dots, d$ there exists an L^1 function g_i such that $|y_i|p(y) \leq g_i(y_i)$ for all $y \in \mathbb{R}^d$.

In addition, for any $d \geq 2$, letting

$$(18) \quad x^{-i} = (x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_d),$$

Assumption 2.2 also holds when there exists some positive δ such that the supports of ν and ν^i , $i = 1, \dots, d$ have empty intersection with a ball around the origin of radius δ , or if the support S of ν satisfies

$$(19) \quad S \subset \bigcap_{i=1, \dots, d} \{x : \|x^{-i}\|_\infty \geq 2\delta\} = \{x : \exists k \neq l \mid |x_k| \geq 2\delta, |x_l| \geq 2\delta\},$$

that is, if every element of the support has at least two nonzero coordinates whose absolute values are larger than some (uniform) constant.

EXAMPLE 2.3. When the distribution of Y has support S_Y , then (19) is satisfied for the support S of $X = Y + \theta$ whenever, for some $\delta > 0$,

$$\theta \in \bigcap_{i=1}^d \{ \psi : \| \psi^{-i} + y^{-i} \|_\infty \geq 2\delta \, \forall y \in S_Y \}.$$

When S_Y is finite, the collection of shifts θ that are excluded can be written as a finite union of sets whose probability with respect to any given absolutely continuous probability measure can be made arbitrarily small by choice of δ .

EXAMPLE 2.4. To see the importance of conditions, such as (19), that require the supports of ν^i to be bounded away from zero, let Y have the uniform distribution on $\{(-1, 0), (1, 0), (0, -1), (0, 1)\}$. Then $Y^1 =_d (U, 0)$ where $U \sim \mathcal{U}[-1, 1]$, and specializing identity (14) to the case $f = (g_{0,1}, 0)$ with g_0 as in (5), we obtain

$$\begin{aligned} E[\sigma_1^2 \partial_1 g_{0,1}(Y^1)] &= E[Y_1 g_{0,1}(Y)], \quad \text{which produces} \\ \frac{1}{4} \int_{-1}^1 \partial_1 g_{0,1}(u, 0) du &= \frac{1}{4} (g_{0,1}(1, 0) - g_{0,1}(-1, 0)). \end{aligned}$$

But the latter identity cannot hold since the left-hand side, with the integral over $[-1, 1]$ of $-1/u^2$ being infinite, and the right-hand side taking the value $1/2$. This example demonstrates that Assumption 2.2 may not hold when the support of ν^i includes zero for some i .

2.3. *Models.* We end this section with a discussion of the various models and to provide some context for the assumptions made for the observation error Y .

MODEL 2.1 (Log-concave). With ϕ denoting a convex function, a (nondegenerate) vector Y is log-concave when it has a density of the form $\exp(-\phi)$ with respect to the Lebesgue measure; ϕ is called the potential. The vector Y is strictly log-concave if the function ϕ is strictly convex, and strongly log-concave when ϕ is strongly convex. When the potential ϕ is two times differentiable, strong log-concavity amounts to the assumption that $\text{Hess}(\phi)(x) \geq \sigma^{-2} \text{Id}$ for all $x \in \mathbb{R}^d$ and for some $\sigma^2 > 0$. Another characterization of strong log-concavity corresponds to assuming the existence of a density $p = \exp(-\psi) \gamma_{\sigma^2}$, where ψ is a convex function and γ_{σ^2} is the density of a Gaussian vector with mean zero and covariance $\sigma^2 \text{Id}$ ([39]).

This class of measures generalizes uniform measures over convex sets to a nonuniform setting. They include of course Gaussian measures, but also many other examples, such as exponential, logistic and Gamma distributions. They are a common class of distributions where one can seek to generalize properties of Gaussian distribution, in particular in high-dimensional settings. They play a role in several areas of applied mathematics, such as convex optimization and optimal transport theory. We refer to [39] for a survey, and to [44] for a panorama of applications. Our work here shall use recent developments in the study of high-dimensional log-concave measures, such as the recent (almost) solution to the KLS conjecture [11, 31], to prove that shrinkage estimation works nicely in this setting.

MODEL 2.2 (Unconditional). Condition (15) shares a strong formal similarity with the notion of a martingale difference random field [34], although in the latter case, the indices of the collection of random variables live in a lattice. A main class of examples of martingale difference random fields depend on “superparity potentials” [34], Definition 3.

Superparity is related to the term “unconditional” that appears in literature linked to high-dimensional geometry (see, e.g., [5]). We say the vector $Y = (Y_1, \dots, Y_d)$ is *unconditional* when it satisfies

$$(\epsilon_1 Y_1, \dots, \epsilon_d Y_d) =_d (Y_1, \dots, Y_d) \quad \text{for all } (\epsilon_1, \dots, \epsilon_d) \in \{-1, 1\}^d;$$

such vectors are easily seen to have a diagonal covariance matrix and to satisfy (15).

Zero biasing of unconditional vectors was considered in [23] and (17) recovers a construction used there for this special case. Such vectors arise naturally as the uniform distribution over bodies in \mathbb{R}^d that have a high degree of symmetry, and include spherically-symmetric distributions. However, equation (15) can hold for many nonelliptical examples. Up to an affine transformation, an elliptical distribution has a density of the form $p(\|x\|_2)$, while subject only to the existence of the necessary conditional expectations, equation (15) holds for *any* distribution with density of the form $p(|x_1|, \dots, |x_d|)$, for p with argument in \mathbb{R}_+^d instead of just \mathbb{R}_+ . In particular, there is no need to have an underlying ℓ^2 structure. An already relevant class of measures are those with a density that is a function of some ℓ^p norm. These distributions are elliptical if and only if $p = 2$, while Condition (15) applies to any value of p , including $p = \infty$.

Furthermore, Poincaré inequalities for unconditional distributions have been investigated, for example, in [28], which proved, under an additional assumption of log-concavity, that the Poincaré constant scales at most logarithmically in the dimension; this result allows one to bound the squared Stein discrepancy $E[\|T - \Sigma\|^2]$ in the kernel approach (see also [8], Proposition 2.21, for a simplified proof); this quantity has been used as a measure of how far away a given distribution is from the Gaussian [30, 35], typically in the case $\Sigma = \text{Id}$. A further relevant feature worthy of note here is that if we assume in addition that the distribution of Y is strictly log-concave, then the maximum likelihood estimator of θ is X , as in the Gaussian case.

MODEL 2.3 (Mixture). The most basic multidimensional model for a mean zero random vector Y that satisfies the conditional expectation condition (15) is the one that assumes independence among coordinates; for this model, a Stein kernel was discussed above Remark 2.1. For zero biasing, taking Y to have a nonsingular covariance matrix to exclude trivial cases, one may easily verify that the vectors

$$(20) \quad Y^i = (Y_1, \dots, Y_{i-1}, Y_i^*, Y_{i+1}, \dots, Y_d) \quad \text{for } i = 1, \dots, d$$

satisfy (14), where Y_i^* is independent of $\{Y_j, j \neq i\}$, and has the Y_i -zero biased distribution. In particular, Y and Y^i may be put on a joint space by specifying that Y_i^* is independent of $Y_j, j \neq i$, and fixing any coupling for (Y_i, Y_i^*) .

The independent model can be extended through the use of mixtures, as follows. Let (\mathcal{S}, Σ) be a measurable space and let $\{m_s\}_{s \in \mathcal{S}}$ be a collection of probability measures on \mathbb{R}^d such that for each Borel subset $A \subset \mathbb{R}^d$ the function $s \rightarrow m_s(A)$ from \mathcal{S} to $[0, 1]$ is measurable. When μ is a probability (mixing) measure on (\mathcal{S}, Σ) , the set function given by

$$m_\mu(A) = \int_{\mathcal{S}} m_s(A) \mu(ds)$$

is a probability measure, called the μ mixture of $\{m_s\}_{s \in \mathcal{S}}$. We may also refer to this distribution as the μ mixture of the distributions of random variables $X_s \sim m_s, s \in \mathcal{S}$, and write the equivalent identity

$$(21) \quad E[f(X)] = \int_{\mathcal{S}} E_s[f(X_s)] d\mu$$

for real valued, bounded continuous functions f on \mathbb{R}^d , and E_s denoting expectation with respect to m_s .

Using such mixtures, we extend the basic independent coordinate model to the case where the vector X is of the form $Y + \theta$ for some unknown $\theta \in \mathbb{R}^d$ and Y is the μ mixture of the mean zero distributions $Y_s, s \in \mathcal{S}$ with nonsingular covariance matrices $\Sigma_s, s \in \mathcal{S}$, each satisfying (15), and where $\Sigma = \text{Var}(Y)$ is finite. This extension gives rise to a wide collection of distributions that generally have dependent coordinates. It also handles the case where a Gaussian observation has positive probability of being corrupted by noise; see Example 3.5.

In particular, under this model, condition (15) holds for Y , and in the special case where Σ_s does not depend on $s \in \mathcal{S}$ then Proposition 5.1 shows that for $i = 1, \dots, d$, the distribution of the zero bias vector Y^i exists, and is simply the μ mixture of Y_s^i . Included are cases where for all $s \in \mathcal{S}$ the d components of Y_s are independent with variance not depending on s , in which case the zero bias vectors for the distributions in the mixture can be constructed as in (20).

It is possible to build a Stein kernel for mixtures of centered measures that each have a Stein kernel. In particular, when $T_s, s \in \mathcal{S}$ is a Stein kernel for Y_s , then it is easy to see that T is a Stein kernel for Y when (Y, T) is the μ mixture of (Y_s, T_s) . However, here we shall mostly discuss mixtures in the context of zero-bias transforms.

3. Shrinkage for non-Gaussian models. Consider the shrinkage estimator

(22)
$$S_\lambda(X) = X \left(1 - \frac{\lambda}{\|X\|^2}\right), \quad \lambda \geq 0$$

of an unknown mean θ of a random vector $X \in \mathbb{R}^d$. We have $S_\lambda(X) = X + f(X)$, dropping the dependence of f on λ . Using (12),

(23)
$$f(x) = -\lambda \frac{x}{\|x\|^2} \quad \text{satisfies} \quad \nabla f(x) = -\lambda \left(\frac{1}{\|x\|^2} \text{Id} - \frac{2}{\|x\|^4} x x^\top \right).$$

To explore the mean squared error of $S_\lambda(X)$, when the corresponding expectations exist, expansion yields

(24)
$$E_\theta \|S_\lambda(X) - \theta\|^2 = E_\theta \left\{ \|X - \theta\|^2 - 2\lambda \left\langle X - \theta, \frac{X}{\|X\|^2} \right\rangle + \frac{\lambda^2}{\|X\|^2} \right\},$$

where we now emphasize the goal of estimating the unknown mean θ of X by including it as a subscript. The mean squared error of $S_0(X)$ is given by the first term. Thus $S_\lambda(X)$ has smaller mean squared error than $S_0(X)$ if and only if the sum of the expectations of the two last terms is negative. We apply the two extensions of the Stein identity in Section 2 to reformulate the expectation $E_\theta \langle X - \theta, f(X) \rangle$ of the second term, namely, using the Stein kernel identity (9), and the multidimensional zero-bias transform identity (16). These two approaches yield qualitatively similar results, though they depend on different properties of the X distribution. Exploring both approaches, we show how the use of S_λ provides advantages for estimating the mean in some non-Gaussian settings.

3.1. Stein kernels. In this subsection, we present three theorems that offer generalizations of the Gaussian case, under different assumptions on the Stein kernels of the distribution ν . In Theorem 3.1, it is assumed that ν admits a Stein kernel that is uniformly bounded from above and below. Theorem 3.2 holds for general distributions with positive definite covariance matrices, while Theorem 3.3 addresses distributions for which a Poincaré inequality holds. The proofs of these results are deferred to the end of this subsection.

Considering nonisotropic random vectors, Theorem 3.1 that follows offers a nonparametric generalization of the Gaussian case, and does not require Assumption 2.1.

THEOREM 3.1. *Consider the measure ν of a random vector $X - \theta$ with mean zero such that $E_\theta[\|X\|^{-2}] < \infty$. Assume that ν admits a Stein kernel T that is uniformly bounded from below and above, in the sense that $\alpha_- \text{Id} \leq T \leq \alpha_+ \text{Id}$ a.s. for the usual partial ordering of symmetric matrices, with α_- and α_+ positive constants. Then*

$$(25) \quad E_\theta \|S_\lambda(X) - \theta\|^2 \leq E\|X - \theta\|^2 - \lambda E_\theta \left[\frac{1}{\|X\|^2} \right] (2d\alpha_- - 4\alpha_+ - \lambda),$$

and if $d \geq 1 + \lfloor 2\alpha_+/\alpha_- \rfloor$ and $\lambda \in (0, 2d\alpha_- - 4\alpha_+)$, then the shrinkage estimator S_λ has a smaller risk than the least-squares estimator S_0 . In particular, if ν is the measure of a strongly log-concave random vector $X - \theta$ with full support, mean zero and twice differentiable potential ϕ satisfying $c_+ \text{Id} \geq \text{Hess}(\phi)(x) \geq c_- \text{Id}$ for any $x \in \mathbb{R}^d$ for some constants $c_+, c_- > 0$, then $E_\theta[\|X\|^{-2}] < \infty$ for $d \geq 3$ and (25) holds with $\alpha_- = 1/c_+$ and $\alpha_+ = 1/c_-$.

This result, which proof is detailed in Supplement B, has three main features. First, the classical result for $d \geq 3$ and $X - \theta$ having a normal distribution is recovered from (25) with $\alpha_- = \alpha_+ = 1$ (or $c_+ = c_- = 1$ in the strongly log-concave case), since in this case one can take $T = \Sigma = \text{Id}$. Second, no condition is needed concerning the behavior of $\|\theta\|$ with respect to dimension d . Third, the result in the strongly log-concave case shows that, for sufficiently large dimensions, there exist shrinkage estimators that are asymptotically better than the least-squares estimator (which is also the MLE for an unconditional strictly log-concave noise vector), even in nonisotropic situations and without knowledge of the covariance matrix, or of any need to estimate it. Moreover, we do not require any form of symmetry, unlike previous results on elliptical distributions.

The next result, Theorem 3.2, is much more general than Theorem 3.1, and demonstrates that shrinkage can improve the MSE by providing a bound involving a term that depends on λ that can be negative, plus a term $2B_\lambda$ that measures the discrepancy between the given distribution and the Gaussian, which will be of smaller order under certain conditions. We formalize some conditions under which shrinkage is to advantage in the following remark.

REMARK 3.1. For X with mean θ and covariance matrix Σ , the risk of the estimator S_0 is $\text{Tr}(\Sigma)$, which is the first term in the bound (30) below, while the second term becomes

$$(26) \quad -E_\theta \left[\frac{(\text{Tr}(\Sigma) - 2\kappa)^2}{\|X\|^2} \right] \quad \text{when } \lambda = \text{Tr}(\Sigma) - 2\kappa.$$

When

$$(27) \quad \|\theta\|^2 = O(\text{Tr}(\Sigma)) \quad \text{and} \quad \kappa = o(\text{Tr}(\Sigma))$$

then using Jensen's inequality to obtain

$$(28) \quad E_\theta \left[\frac{1}{\|X\|^2} \right] \geq \frac{1}{\|\theta\|^2 + \text{Tr}(\Sigma)},$$

we see that (26) is negative with absolute value at least on the order of $\text{Tr}(\Sigma)$. Hence, shrinkage with this value of λ will improve the mean squared error over that of S_0 when $B_\lambda = o(\text{Tr}(\Sigma))$. Similar remarks apply in general when $\lambda/2(\text{Tr}(\Sigma) - 2\kappa)$ is bounded away from 0 and 1. In the canonical case where growth is of order d , the set of conditions

$$(29) \quad \|\theta\|^2 + \text{Tr}(\Sigma) = O(d) \quad \text{and} \quad d = O(\text{Tr}(\Sigma) - 2\kappa)$$

are equivalent to the two in (27) along with the additional assumption that $\text{Tr}(\Sigma)/d$ is bounded away from zero and infinity.

THEOREM 3.2. *Let a random vector X have mean θ , positive semidefinite covariance matrix Σ with largest eigenvalue κ , and Stein kernel $T = T_{X-\theta}$. Then*

$$(30) \quad E_{\theta} \|S_{\lambda}(X) - \theta\|^2 \leq E \|X - \theta\|^2 + E_{\theta} \left[\frac{\lambda}{\|X\|^2} (\lambda - 2(\text{Tr}(\Sigma) - 2\kappa)) \right] + 2B_{\lambda},$$

where

$$(31) \quad B_{\lambda} = |E_{\theta}[\langle \Sigma - T_{X-\theta}, \nabla f(X) \rangle]|$$

with f as in (23), and

$$(32) \quad B_{\lambda} \leq \frac{\lambda}{d} \sqrt{E_{\theta}[d^2 \|X\|^{-4}]} \{ \sqrt{\text{Var}(\text{Tr}(T))} + 2\sqrt{E[\|T - \Sigma\|^2]} \}.$$

If for some d_0 ,

$$(33) \quad \sup_{d \geq d_0} E_{\theta}[d^2 \|X\|^{-4}] < \infty, \quad \text{Var}(\text{Tr}(T)) = o(d^2) \quad \text{and} \quad E[\|T - \Sigma\|^2] = o(d^2)$$

and $\lambda = O(d)$, then $B_{\lambda} = o(d)$, and over the range $\lambda \in [0, 2(\text{Tr}(\Sigma) - 2\kappa)]$ the risk of S_{λ} is no larger than that of S_0 asymptotically. If in addition $\lambda/2(\text{Tr}(\Sigma) - 2\kappa)$ is bounded away from zero and (29) holds then the shrinkage estimator (22) has strictly smaller mean squared error than $S_0(X) = X$ for all d sufficiently large.

The proof of Theorem 3.2 can be found in Supplement B. The quantity $\text{Tr}(\Sigma) - 2\kappa$ in (30) is used to bound the trace of a matrix product that appears in the proof, and could be refined at the cost of an expression involving higher moments. In the special case when $\Sigma = \sigma^2 \text{Id}$, (30) simplifies to

$$(34) \quad E_{\theta} \|S_{\lambda}(X) - \theta\|^2 \leq E \|X - \theta\|^2 + E_{\theta} \left[\frac{\lambda}{\|X\|^2} (\lambda - 2\sigma^2(d - 2)) \right] + 2B_{\lambda}.$$

For ease of notation, the dependence of the bound B_{λ} on the negative moments of $\|X\|^2$ and the choice of kernel T is suppressed.

One condition needed for some of the results that follow is that for some $m \geq 1$ there exists a constant C such that

$$(35) \quad (a) \quad E_{\theta} \left[\frac{d}{\|X\|^2} \right]^m \leq C \quad \text{or} \quad (b) \quad \max_{1 \leq i \leq d} E_{\theta} \left(\frac{d}{\|X^{-i}\|^2} \right)^m \leq C,$$

where we recall the notation in (18) for part (b). These conditions are handled in Section 6, and shown to be satisfied, for instance, under log-concavity in Proposition 6.1, the technical condition (77) given in Lemma 6.1 and discussed in Remark 6.1.

Lemma 2.1, that gives conditions under which Assumption 2.1 holds, requires us to work in dimension at least 5, while the critical dimension in the Gaussian case is just 3. Under suitable integrability conditions, one would expect the critical dimension implicit in Theorem 3.2 to decrease to 3 as the sample size n in Example 3.1 tends to infinity. However, to ensure that the shrinkage estimator is allowed in the Stein identity with only the required weaker integrability condition, one would have to only use L^1 estimates on its gradient, instead of L^2 , which by duality would require us to work with L^{∞} Stein kernels, as in 3.1. But we do not expect bounded Stein kernels to exist for simply log-concave distributions, for example, and so in that more general setting we shall assume $d \geq 5$.

Applying Theorem 3.2 to Model 2.3 we see how shrinkage may be to advantage in non-Gaussian situations. A short proof of Corollary 3.1 is given in Supplement B, illustrating how the conditions of Theorem 3.2 can be easily verified.

COROLLARY 3.1. Suppose that $\mathbf{X} \in \mathbb{R}^d$ satisfies the conditions of Model 2.3, where the components $Y_{s,i}$ of \mathbf{Y}_s are independent and $\Sigma_s = \sigma^2 \text{Id}$ for all $s \in S$, the Stein kernels $T_{s,i}$ of the components \mathbf{Y}_s satisfy $\sup_{s \in S, 1 \leq i \leq d} E[T_{s,i}^2] < \infty$, and that for all $s \in S$, (35) (part (a)) with $d = 2$ is satisfied by $\|\mathbf{X}_s\| = \|\mathbf{Y}_s + \boldsymbol{\theta}\|$. Then if $\lambda = \sigma^2(d - 2)$ and $\|\boldsymbol{\theta}\|^2 = O(d)$, the shrinkage estimator (22) has strictly smaller mean squared error than S_0 for all d sufficiently large.

EXAMPLE 3.1. If $T_{X_i - \boldsymbol{\theta}}, i = 1, \dots, n$ are Stein kernels for an independent sample of vectors $\mathbf{X}_i, i = 1, \dots, n$, each with mean $\boldsymbol{\theta}$ and covariance matrix $\sigma^2 \text{Id}$, then as noted in [6] in one dimension, a Stein kernel T for their average

$$\mathbf{X} = \frac{1}{n} \sum_{i=1}^n \mathbf{X}_i \quad \text{is given by} \quad T = \frac{1}{n} \sum_{i=1}^n T_{X_i - \boldsymbol{\theta}},$$

since, by independence,

$$\begin{aligned} E_{\boldsymbol{\theta}}[\langle \mathbf{X} - \boldsymbol{\theta}, f(\mathbf{X}) \rangle] &= \left[\frac{1}{n} \sum_{i=1}^n E_{\boldsymbol{\theta}}[\langle \mathbf{X}_i - \boldsymbol{\theta}, f(\mathbf{X}) \rangle] \right] \\ &= \frac{1}{n} \sum_{i=1}^n E_{\boldsymbol{\theta}}[\langle T_{X_i - \boldsymbol{\theta}}, \nabla f(\mathbf{X}) \rangle] = E_{\boldsymbol{\theta}}[\langle T, \nabla f(\mathbf{X}) \rangle]. \end{aligned}$$

Under conditions on the measures of the average, results of [13] guarantee existence and uniqueness of Stein kernels within $W_{v,0}^{1,2}$, the set of functions in $W^{1,2}(v)$ with v -mean zero. As $E[T_i] = \text{Var}(\mathbf{X}_i)$ via (9), we see that $\text{Var}(\text{Tr}(T))$ and $E\|\mathbf{T} - \sigma^2 \text{Id}\|^2$, and hence the quantities in the last two conditions of (3.2) in Theorem 3.2, will decrease in n under mild moment conditions.

Next, we illustrate Theorem 3.2 using an explicit example of a random vector having dependent coordinates, allowing us to obtain precise results for this particular case.

EXAMPLE 3.2. Consider $\mathbf{X} = \mathbf{Y} + \boldsymbol{\theta}$ in \mathbb{R}^d with \mathbf{Y} from the family of multivariate central Student- t distributions, taken here with $k \geq 5$ degrees of freedom, shape given by a symmetric, positive definite matrix Υ in $\mathbb{R}^{d \times d}$ and $d = 2m \geq 6$, even. These distributions are the subfamily of the elliptical distributions introduced in Example 2.1, obtained by taking

$$(36) \quad \phi(t) = (1 + 2t/k)^{-(k+d)/2};$$

the covariance matrix of \mathbf{Y} is $\Sigma = (k/(k-2))\Upsilon$. Here, we take $\Upsilon = \text{Id}$ so that $\Sigma = \sigma^2 \text{Id}$ for $\sigma^2 = k/(k-2)$. Using that $d + k > 2$, from (36) followed by (11), we obtain that

$$(37) \quad \frac{1}{\phi(t/2)} \int_{t/2}^{+\infty} \phi(u) du = \frac{t+k}{d+k-2} \quad \text{and hence} \quad T = \left(\frac{\mathbf{Y}^\top \mathbf{Y} + k\sigma^2}{d+k-2} \right) \text{Id}$$

is a Stein kernel for the multivariate Student distribution; see also [33] and [29]. We obtain the following bounds for the terms controlling B_λ in the right-hand side of inequality (32) of Theorem 3.2 (see Supplement B for details):

$$E_{\boldsymbol{\theta}}[d^2 \|\mathbf{X}\|^{-4}] \leq \frac{d^2(k-2)^2(k+2)}{(d-2)(d-4)k^3}, \quad \text{Var}(\text{Tr}(T)) = \frac{2d^3k^4}{(d+k-2)(k-2)^4(k-4)}$$

and

$$E[\|\mathbf{T} - \sigma^2 \text{Id}\|^2] = \frac{2d^2k^4}{(d+k-2)(k-2)^4(k-4)}.$$

Both these last two terms are $o(d^2)$ as long as $1/k = o(1)$, in which case all conditions in (33) hold.

We note that in view of Example 2.1, Example 3.2 can be translated into a Student distribution with k degrees of freedom, with any parameter Σ , having covariance matrix $(k/(k - 2))\Sigma$.

A particular instance of Theorem 3.2 arises when the (centered) distribution ν with finite covariance matrix Σ satisfies a Poincaré inequality

(38)
$$\mathrm{Var}_\nu(f) \leq C_P E_\nu[\|\nabla f(\mathbf{X})\|^2]$$

for all functions f for which the quantities make sense and the right-hand side is finite. In [13], using techniques that already appeared in [43], under Assumption 3.1 it is shown (see equation (6) *ibid.*) that in this situation a Stein kernel T for ν exists that satisfies

$$E_\nu[\|T\|^2] \leq C_P \mathrm{Tr}(\Sigma).$$

Some caveats are addressed through Assumption 3.1 below. Hence, using $E_\nu[\mathrm{Tr}(T)] = \mathrm{Tr}(\Sigma) \leq \kappa d$ and $\mathrm{Tr}(\Sigma^2) \geq (\mathrm{Tr}(\Sigma))^2/\mathrm{rank}(\Sigma)$, we obtain

(39)
$$E_\nu[\|T - \Sigma\|^2] = E_\nu[\|T\|^2 - \mathrm{Tr}(\Sigma^2)] \leq \kappa d \left(C_P - \frac{\mathrm{Tr}(\Sigma)}{\mathrm{rank}(\Sigma)} \right).$$

In particular, if the Poincaré constant is independent of the dimension (which is the case for product measures), then the (squared) Stein discrepancy on the left-hand side is of order d , which is negligible compared to d^2 , so that the third requirement of Theorem 3.2 will be satisfied. Another family of examples that can be included using these methods are log-concave distributions; see Corollary 3.2 below.

Here, we clarify some issues related to the space of admissible functions for the Stein identity proved in [13]. The argument there proves existence of a Stein kernel such that the identity (9) holds for functions lying in the closure of smooth, compactly supported test functions with respect to the Sobolev norm. This closure may not be the whole space $W^{1,2}(\nu)$. To bypass this issue, we make the following extra assumptions, which in full generality is stronger than Assumption 2.1, but equivalent in most situations (such as when ν has a continuous density with full support, or a support with smooth boundary and density bounded away from zero [18], Section 5.3.3).

ASSUMPTION 3.1. The function $\mathbf{g}_0(\mathbf{x}) = \mathbf{x}/\|\mathbf{x}\|^2$ is in the closure of $C_c^\infty(\mathbb{R}^d)$, the set of infinitely differentiable functions with compact support, with respect to the Sobolev norm (2).

While a bound on the squared Stein discrepancy $E[\|T - \Sigma\|^2]$ is not enough to get suitable control on the variance of $\mathrm{Tr}(T)$, the conclusion of Theorem 3.2 continues to hold if we assume a Poincaré inequality with a constant growing sufficiently slowly, and a stronger moment assumption.

THEOREM 3.3. Let Assumption 3.1 be satisfied for the measure ν of a random vector \mathbf{X} with mean $\boldsymbol{\theta}$, covariance matrix Σ with largest eigenvalue κ , and for which the Poincaré inequality holds with constant C_P as given in (38). If for some d_0 , we have

(40)
$$\sup_{d \geq d_0} E_\theta[d^3 \|\mathbf{X}\|^{-6}] < \infty \quad \text{and} \quad C_P = o(\sqrt{d}),$$

$\lambda = O(d)$, then $B_\lambda = o(d)$ and when $\lambda \in [0, 2(\mathrm{Tr}(\Sigma) - 2\kappa)]$ the risk of S_λ is no larger than that of S_0 asymptotically, and if $\lambda/(2(\mathrm{Tr}(\Sigma) - 2\kappa))$ is bounded away from zero and one as d tends to infinity and the conditions in (29) hold, the shrinkage estimator (22) has strictly smaller mean squared error than $S_0(\mathbf{X}) = \mathbf{X}$ for all d sufficiently large.

This latter result, whose proof is given in Supplement B, contains the case of a vector with independent coordinates and finite Poincaré constant, since that is then dimension-free. Another family of examples is given by the following corollary, which can be applied to certain isotropic random vectors.

COROLLARY 3.2. *For any $A > 0$, there exists a critical dimension d_0 that only depends on A such that if $d \geq d_0$, then for any measure ν of an isotropic log-concave random vector $\mathbf{X} - \boldsymbol{\theta}$ with mean zero and covariance matrix Id that satisfies Assumption 3.1, the risk of S_λ is strictly smaller than the risk of S_0 for $\lambda = d - 2$, corresponding to the classical James–Stein estimator, as long as $\|\boldsymbol{\theta}\|^2 \leq Ad$.*

Considering nonisotropic random vectors, the following corollary also holds and still offers a nonparametric generalization of the Gaussian case.

COROLLARY 3.3. *Consider the measure ν of a strongly log-concave random vector $\mathbf{X} - \boldsymbol{\theta}$ with mean zero, covariance matrix Σ and two times differentiable potential ϕ . Assume that there are constants $c, l > 0$ such that in the partial order on symmetric matrices $\text{Hess}(\phi)(\mathbf{x}) \geq c \text{Id}$, for any $\mathbf{x} \in \mathbb{R}^d$, and that the smallest eigenvalue of the covariance matrix is greater than l . Then there exists a critical dimension d_0 , such that if $d \geq d_0$, the risk of S_λ is strictly smaller than the risk of S_0 for $\lambda = O(d)$, as long as $\|\boldsymbol{\theta}\|^2 = O(d)$.*

The proof of Corollaries 3.2 and 3.3 can be found in Supplement B. These results show that there exist shrinkage estimators that are asymptotically better than the MLE (which is indeed S_0 for an unconditional strictly log-concave noise vector), even in nonisotropic situations and without estimating the covariance matrix. However, taking into account the covariance structure in the estimator may lead to better performances. We leave this question for future work.

Before beginning the proofs of Theorems 3.1, 3.2 and 3.3, we note that by using the form of the Jacobian (23), for any nonnegative definite matrix M with largest eigenvalue bounded by κ , and $\mathbf{f} \in W_{1,2}(\nu)$,

$$\begin{aligned} E_{\boldsymbol{\theta}}[\langle M, \nabla \mathbf{f} \rangle] &= -\lambda E_{\boldsymbol{\theta}} \left[\left\langle M, \frac{1}{\|\mathbf{X}\|^2} \text{Id} - \frac{2}{\|\mathbf{X}\|^4} \mathbf{X} \mathbf{X}^\top \right\rangle \right] \\ (41) \quad &= -\lambda E_{\boldsymbol{\theta}} \left[\frac{\text{Tr}(M)}{\|\mathbf{X}\|^2} - \frac{2 \text{Tr}(M \mathbf{X} \mathbf{X}^\top)}{\|\mathbf{X}\|^4} \right] \leq -\lambda E_{\boldsymbol{\theta}} \left[\frac{\text{Tr}(M) - 2\kappa}{\|\mathbf{X}\|^2} \right], \end{aligned}$$

where we used that the trace of a product of two nonnegative definite matrices can be bounded the largest eigenvalue of one multiplied by the trace of the other.

3.2. Zero bias. We now derive analogous shrinkage results based on zero biasing. We will write the Euclidean norm as $\|\mathbf{x}\|_2$ when it appears in proximity to other p norms. In the following, we will take advantage of the fact that differences of an expression involving only \mathbf{X} with one involving only \mathbf{X}^{ij} depend only on the (marginal) distributions of \mathbf{X} and \mathbf{X}^{ij} , and hence not on the choice of coupling. For this reason, in expressions such as (43) we may assume that these vectors are jointly given without specifying a joint distribution. Conditions under which the zero bias vectors required in the following result exist are given in Proposition 5.1.

THEOREM 3.4. *Let $\mathbf{X} = \mathbf{Y} + \boldsymbol{\theta}$ where $\mathbf{Y} \in \mathbb{R}^d$ has covariance matrix Σ with largest eigenvalue κ , and suppose that for all pairs i, j such that $\sigma_{ij} \neq 0$ the zero bias vectors \mathbf{X}^{ij}*

exist as in (16). Then

$$(42) \quad E_{\theta} \|S_{\lambda}(\mathbf{X}) - \theta\|^2 \leq E_{\theta} \|\mathbf{X} - \theta\|^2 + E_{\theta} \left[\frac{\lambda}{\|\mathbf{X}\|^2} (\lambda - 2(\text{Tr}(\Sigma) - 2\kappa)) \right] + 2B_{\lambda}^*,$$

where, with f as in (5),

$$(43) \quad B_{\lambda}^* = \left| E_{\theta} \sum_{i,j=1}^d \sigma_{ij} [\partial_j f_i(\mathbf{X}^{ij}) - \partial_j f_i(\mathbf{X})] \right|.$$

When $\Sigma = \sigma^2 \text{Id}$,

$$(44) \quad B_{\lambda}^* = \lambda \sigma^2 \left| E_{\theta} \sum_{i=1}^d \left(\frac{\|\mathbf{X}^i\|^2 - 2(X_i^i)^2}{\|\mathbf{X}^i\|^4} \right) - \left(\frac{\|\mathbf{X}\|^2 - 2X_i^2}{\|\mathbf{X}\|^4} \right) \right|.$$

Assuming $\lambda \in [0, 2\sigma^2(d-2)]$ and that $\|\theta\|^2 = O(d)$, the following two scenarios lead to simplified bounds:

1. Let \mathbf{X} satisfy (35) (part b) with $m = 4$ and constant C_{-4} , and have components that for $r = 4$ and 8 satisfy $\sup_{i=1,2,\dots,d} E(X_i - \theta_i)^r \leq C_r$. Then with notation as in (18), for all $d \geq 3$,

$$(45) \quad \begin{aligned} B_{\lambda}^* &\leq \lambda \sum_{i=1}^d |\text{Cov}_{\theta}((X_i - \theta_i)^2, \|\mathbf{X}^{-i}\|^{-2})| \\ &\quad + \frac{6\lambda\sqrt{C_{-4}}}{d^2} (d(\sigma^2\sqrt{C_4} + \sqrt{C_8}/3) + \|\theta\|^2(\sigma^2 + C_4)). \end{aligned}$$

When $\sum_{i=1}^d |\text{Cov}_{\theta}((X_i - \theta_i)^2, \|\mathbf{X}^{-i}\|^{-2})| = o(1)$ then $B_{\lambda}^* = o(d)$, and when the ratio $\lambda/(2\sigma^2(d-2))$ stays bounded away from 0 and 1 the shrinkage estimator has strictly smaller mean squared error than S_0 for all d sufficiently large.

2. Let \mathbf{X} follow Model 2.3, where \mathbf{Y} is the μ mixture of \mathbf{Y}_s , $s \in \mathcal{S}$, with \mathbf{Y}_s having mean zero, covariance matrix $\sigma^2 \text{Id}$ and independent components, and let (35) (part b) hold for $\mathbf{X}_s = \mathbf{Y}_s + \theta$ for $m = 2$ for all $s \in \mathcal{S}$ and $i = 1, 2, \dots, d$ with constant C_{-2} . Then for any $X_{i,s}^*$ on a joint space with \mathbf{X}_s , independent of \mathbf{X}_s^{-i} and having the $X_{i,s}$ -zero bias distribution,

$$(46) \quad B_{\lambda}^* \leq \frac{25C_{-2}\lambda\sigma^2}{8d^2} \sum_{i=1}^d \int_{\mathcal{S}} E_{\theta,s} |X_{s,i}^2 - (X_{s,i}^*)^2| d\mu.$$

If $\max_{1 \leq i \leq d} E[Y_{s,i}^4] \leq C_4$ for all $s \in \mathcal{S}$, then

$$(47) \quad B_{\lambda}^* \leq \frac{25C_{-2}\lambda}{8d^2} (dC_4/3 + C_4^{3/4}\|\theta\|_1 + 2\sigma^2\|\theta\|_2^2 + d\sigma^4),$$

and $B_{\lambda}^* = O(1)$. When the ratio $\lambda/(2\sigma^2(d-2))$ stays bounded away from 0 and 1 the shrinkage estimator has strictly smaller mean squared error than S_0 for all d sufficiently large.

We note that the bounds (43), (44) and (46) are tight, returning zero when the observation vector is Gaussian. We next present Remark 3.2, discussing the covariance term in the bound in Part 1, followed by Examples 3.3, 3.4 and 3.5. The first two of these examples illustrate the computation of the bounds presented in the main part of the theorem, followed by an application of our results to corrupted Gaussian observations. The proof of Theorem 3.4 is deferred to Supplement B.

REMARK 3.2. In the simplest case, the covariance terms in the sum in the bound (45) in Part 1, Theorem 3.4 will be zero when \mathbf{X} follows Model 2.3 with \mathbf{Y} the mixture of \mathbf{Y}_s , $s \in \mathcal{S}$ having independent components. This term will also be of the desired order $o(1/d)$ when the dependence between the components of \mathbf{X} is sufficiently weak. For example, this order obtains when the components of \mathbf{X} are locally dependent with sufficiently small dependency neighborhoods, and certain technical moment bounds on inverse norms, discussed in Section 6, are in force.

Precisely, say for each $i = 1, \dots, d$ there exists a dependency neighborhood $\{i\} \subset \mathcal{N}_i \subset \{1, \dots, d\}$ of size η such that X_i is independent of $\{X_j, j \notin \mathcal{N}_i\}$. Then, suppressing the dependence on i when defining U and V , we may write

$$\|\mathbf{X}^{-i}\|^2 = U + V \quad \text{where } U = \sum_{j \in \mathcal{N}_i \setminus \{i\}} X_j^2 \quad \text{and} \quad V = \sum_{j \notin \mathcal{N}_i} X_j^2$$

$$\text{and let } W = (X_i - \theta_i)^2 - \sigma^2.$$

Then, using that W has mean zero and is independent of V , and that U and V are nonnegative, applying Hölder's inequality in the final step, we obtain

$$\begin{aligned} & |\text{Cov}_\theta((X_i - \theta_i)^2, \|\mathbf{X}^{-i}\|^{-2})| \\ &= \left| E\left[\frac{W}{U+V}\right] \right| = \left| E\left[\frac{W}{U+V} - \frac{W}{V}\right] \right| = \left| E\left[\frac{WU}{V(U+V)}\right] \right| \\ &\leq E\left[\frac{|W|U}{V(U+V)}\right] \leq E\left[\frac{|W|U}{V^2}\right] \leq E[W^4]^{1/4} E[U^4]^{1/4} E[V^{-4}]^{1/2}. \end{aligned}$$

When the eighth moments of the components of \mathbf{Y} are uniformly bounded by C_8 ,

$$EW^4 \leq 8(C_8 + \sigma^8) \quad \text{and} \quad EU^4 \leq 64\eta^4(C_8 + \|\theta\|_\infty^8).$$

Hence, when V satisfies (35) (part (b)) for $m = 4$ when playing the role of $\|\mathbf{X}^{-i}\|^2$, we obtain

$$|\text{Cov}_\theta((X_i - \theta_i)^2, \|\mathbf{X}^{-i}\|^{-2})| \leq \frac{8\eta[(C_8 + \sigma^8)(C_8 + \|\theta\|_\infty^8)]^{1/4}}{(d - \eta)^2} = o(1/d)$$

as desired, when η and $\eta\|\theta\|_\infty^2$ are both $o(d)$.

EXAMPLE 3.3. We illustrate how coupling can allow for a computation of a bound on (44). Let \mathbf{U} and \mathbf{U}^* be the uniform distributions on S^{d-1} and B^d , the sphere and unit ball in \mathbb{R}^d , respectively. The divergence theorem, and $\text{Area}(S^{d-1})/\text{Vol}(B^d) = d$, yield that for smooth $f: \mathbb{R}^d \rightarrow \mathbb{R}^d$,

$$E[(\mathbf{U}, f(\mathbf{U}))] = \frac{1}{d} E[\nabla \cdot f(\mathbf{U}^*)],$$

and hence $E[\mathbf{U}] = 0$, $\text{Cov}(\mathbf{U}) = \text{Id}/d$ and the zero bias vectors $\mathbf{U}^i =_d \mathbf{U}^*$ for all $i = 1, \dots, d$. For $\sigma > 0$, letting

$$\mathbf{X} = \boldsymbol{\theta} + \sigma\sqrt{d}\mathbf{U} \quad \text{and} \quad \mathbf{X}^* = \boldsymbol{\theta} + \sigma\sqrt{d}\mathbf{U}^*$$

we have $\mathbf{X}^i = \mathbf{X}^*$, $i = 1, \dots, n$, $E[\mathbf{X}] = \boldsymbol{\theta}$ and $\text{Cov}(\mathbf{X}) = \sigma^2 \text{Id}$.

The distribution of $R^\sharp = \|\mathbf{U}^*\|$ is given by

$$P(R^\sharp \leq r) = P(\|\mathbf{U}^*\| \leq r) = \frac{\text{Vol}(rB^d)}{\text{Vol}(B^d)} = r^d \quad \text{for } 0 \leq r \leq 1,$$

and hence has density in the unit interval given by

$$\frac{d}{dr}P(R^\sharp \leq r) = dr^{d-1} \quad \text{and so} \quad E[R^\sharp] = \int_0^1 dr^d dr = \frac{d}{d+1}.$$

Now, letting \mathbf{U} “pick” a uniform direction, and an independent variable R^\sharp with the density above “pick” a relative magnitude, we obtain the coupling

$$\mathbf{X} = \boldsymbol{\theta} + \sigma\sqrt{d}\mathbf{U} \quad \text{and} \quad \mathbf{X}^* = \boldsymbol{\theta} + \sigma\sqrt{d}R^\sharp\mathbf{U},$$

and that

$$(48) \quad E\|\mathbf{U} - \mathbf{U}^*\| = E\|\mathbf{U} - R^\sharp\mathbf{U}\| = E(1 - R^\sharp) = 1 - \frac{d}{d+1} \leq \frac{1}{d}.$$

Hence, for $h: \mathbb{R}^d \rightarrow \mathbb{R}$ and $\alpha = \sup_S \|\nabla h(\mathbf{x})\| < \infty$ for S the union of the supports of \mathbf{X} and \mathbf{X}^* ,

$$(49) \quad E|h(\mathbf{X}) - h(\mathbf{X}^*)| \leq \alpha E\|\mathbf{X} - \mathbf{X}^*\| = \alpha\sigma\sqrt{d}E\|\mathbf{U} - \mathbf{U}^*\| \leq \frac{\alpha\sigma}{\sqrt{d}}.$$

Specializing (44) to the case where $\mathbf{X}_i = \mathbf{X}^*$ for all $i = 1, \dots, d$, now taking $d \geq 3$, we obtain

$$(50) \quad \begin{aligned} B_\lambda^* &= \lambda\sigma^2 \left| E_\theta \sum_{i=1}^d \left(\frac{\|\mathbf{X}^*\|^2 - 2(X_i^*)^2}{\|\mathbf{X}^*\|^4} \right) - \left(\frac{\|\mathbf{X}\|^2 - 2X_i^2}{\|\mathbf{X}\|^4} \right) \right| \\ &= \lambda\sigma^2(d-2) \left| E_\theta \left(\frac{1}{\|\mathbf{X}^*\|^2} - \frac{1}{\|\mathbf{X}\|^2} \right) \right|. \end{aligned}$$

Taking $\|\boldsymbol{\theta}\|^2 \geq c\sigma^2d$ for some $c > 1$, we have $\min(\|\mathbf{X}\|^2, \|\mathbf{X}^*\|^2) \geq (\|\boldsymbol{\theta}\| - \sigma\sqrt{d})^2 \geq (\sqrt{c} - 1)^2\sigma^2d$, and hence, over the supports of \mathbf{X} and \mathbf{X}^* ,

$$\left\| \nabla \frac{1}{\|\mathbf{x}\|^2} \right\| = \left\| \frac{2\mathbf{x}}{\|\mathbf{x}\|^4} \right\| = 2\|\mathbf{x}\|^{-3} \leq \frac{2}{(\sqrt{c} - 1)^3\sigma^3d^{3/2}}.$$

Now, using (50) and (49) with α as the upper bound above, and taking $\lambda = \sigma^2(d-2)$, we obtain

$$(51) \quad 2B_\lambda^* \leq 2\sigma^4(d-2)^2 \times \frac{\alpha\sigma}{\sqrt{d}} = \frac{4\sigma^2(d-2)^2}{(\sqrt{c} - 1)^3d^2}.$$

When $\|\boldsymbol{\theta}\|^2 \leq C\sigma^2d$,

$$\|\mathbf{X}\| = \|\boldsymbol{\theta} + \sigma\sqrt{d}\mathbf{U}\| \leq \|\boldsymbol{\theta}\| + \sigma\sqrt{d} \leq (\sqrt{C} + 1)^2\sigma\sqrt{d},$$

and (26) yields the upper bound on the second term in the bound (42),

$$(52) \quad -\sigma^4(d-2)^2 E_\theta \left[\frac{1}{\|\mathbf{X}\|^2} \right] \leq \frac{-\sigma^2(d-2)^2}{(\sqrt{C} + 1)^2d}.$$

Comparing to (51), we see shrinkage will strictly improve the mean squared error when

$$d > \frac{4(\sqrt{C} + 1)^2}{(\sqrt{c} - 1)^3}.$$

For instance, when $c = 4$ and $C = 9$ shrinkage is advantageous when $d > 64$.

More generally, we illustrate in Supplement B the use of the bound (43) for a case where the observations have a nondiagonal covariance matrix.

EXAMPLE 3.4. For \mathbf{Y} having the mean zero, variance $\sigma^2 \text{Id}$ Student distribution with k degrees of freedom, $d = 2m$ even, an explicit zero bias coupling can be constructed using the representation of the distribution of \mathbf{Y} as $\mathbf{Y}_\gamma = \gamma^{-1/2} \sigma \mathbf{N}$, with $\mathbf{N} \sim \mathcal{N}_d(0, \text{Id})$, mixed over $\gamma \sim \Gamma(k/2, k/2)$, as outlined in Example 3.2. Part 4 of Proposition 5.1 gives that, for $i = 1, \dots, d$, the zero bias vectors \mathbf{Y}^i are given by the mixture \mathbf{Y}_δ where the distribution of δ has a Radon–Nikodym derivative with respect to the distribution of γ equal to $\text{Var}(Y_{\gamma,i})/\text{Var}(Y_\gamma)$, that is, proportional to γ^{-2} , and hence, $\delta \sim \Gamma(k/2 - 1, k/2)$. Now letting $\epsilon \sim \Gamma(1, k/2)$ be independent of δ , with both variables independent of \mathbf{N} , a coupling of γ and δ is achieved by setting $\gamma = \delta + \epsilon$. Hence Part 4 of Proposition 5.1, and the fact that the normal is fixed by the zero bias transformation, yield the couplings

$$(53) \quad \mathbf{X} = \boldsymbol{\theta} + \frac{\sigma}{\sqrt{\delta + \epsilon}} \mathbf{N} \quad \text{and} \quad \mathbf{X}^i = \boldsymbol{\theta} + \frac{\sigma}{\sqrt{\delta}} \mathbf{N}, \quad i = 1, \dots, d.$$

Using the latter coupling to bound the right-hand side of (44), after computations that are detailed in Supplement B, one can get for $\boldsymbol{\theta} = \mathbf{0}$,

$$B_\lambda^* \leq \frac{2\lambda}{k}.$$

This bound is $o(d)$ when $\lambda = O(d)$ and $1/k = o(1)$. In the case where $\boldsymbol{\theta} \neq \mathbf{0}$,

$$B_\lambda^* \leq \frac{8\lambda(d + k - 2)}{(d - 2)k}$$

and if $\lambda = O(d)$ and $1/k = o(1)$, this bound is $o(d)$ as desired.

EXAMPLE 3.5. Let $\mathbf{X} = \mathbf{Y} + \boldsymbol{\theta}$ where \mathbf{Y} is a Gaussian vector $\mathbf{Y}_0 \sim \mathcal{N}_d(0, \sigma^2 \text{Id})$ corrupted by a mean zero, variance $\sigma^2 \text{Id}$ outlier vector \mathbf{Y}_1 that satisfies assumption (15). By Part 1 of Proposition 5.1, the zero bias vectors of \mathbf{Y}_1 exist.

One corruption model is additive, where for some $\epsilon \in [0, 1]$, $\mathbf{Y} = \sqrt{1 - \epsilon} \mathbf{Y}_0 + \sqrt{\epsilon} \mathbf{Y}_1$. By Part 3 of Proposition 5.1, the zero bias vectors of \mathbf{Y} exist and can be coupled to \mathbf{Y} via

$$\mathbf{Y}^i = \begin{cases} \sqrt{1 - \epsilon} \mathbf{Y}_0^i + \sqrt{\epsilon} \mathbf{Y}_1^i \\ \sqrt{1 - \epsilon} \mathbf{Y}_0 + \sqrt{\epsilon} \mathbf{Y}_1^i \end{cases} = \begin{cases} \sqrt{1 - \epsilon} \mathbf{Y}_0 + \sqrt{\epsilon} \mathbf{Y}_1 & \text{with probability } 1 - \epsilon, \\ \sqrt{1 - \epsilon} \mathbf{Y}_0 + \sqrt{\epsilon} \mathbf{Y}_1^i & \text{with probability } \epsilon, \end{cases}$$

where we have used the normality of \mathbf{Y}_0 to replace \mathbf{Y}_0^i by \mathbf{Y}_0 . With \mathbf{Y}_1 additionally satisfying the conditions in Part 2 of Theorem 3.4, the bound (44) holds with the reduction factor of ϵ over its value for \mathbf{Y}_1 .

Another way that the outlier can enter is via mixing, where with probability $1 - \epsilon$ the vector \mathbf{Y} is the Gaussian \mathbf{Y}_0 , and with probability ϵ equals \mathbf{Y}_1 . By Part 4 of Proposition 5.1, the zero bias vector \mathbf{Y}^i is the same $1 - \epsilon, \epsilon$ mixture of $\mathbf{Y}_0^i = \mathbf{Y}_0$ and \mathbf{Y}_1^i . In particular, the bound (44) takes the value zero with probability $1 - \epsilon$ and, therefore, equals ϵ of the value it has for \mathbf{Y}_1 .

In summary, the four main results of this section, Theorems 3.1, 3.2, 3.3 and 3.4, provide bounds for the mean squared error of the shrinkage estimator S_λ , which is shown to be strictly smaller than that of S_0 under a variety of conditions. The first three of these results depend on Stein kernels, and the fourth on zero biasing. The application at hand determines which of the two types of results would be more straightforward to apply. For instance, Theorem 3.2 requires the existence of a Stein kernel T and that it satisfies certain moment conditions, whereas Theorem 3.4 requires that the observation \mathbf{X} itself satisfies (15), as well as additional moment assumptions. Examples 3.2 and 3.4 illustrate that in some situations when both types of results can be applied, they may yield subtly different bounds.

4. SURE: Stein unbiased risk estimate. In this section, we demonstrate that in some settings the bias incurred when using standard forms of SURE as in (4) when the observation X is not Gaussian can be controlled, and is sufficiently small so as to yield estimates useful for the selection of tuning parameters.

We start by reviewing the Gaussian case for SURE. Though the classical is where the covariance of the observation is $\sigma^2 \text{Id}$ with known σ^2 , we continue with the more general instance where the covariance Σ is a known matrix in order to illustrate the range of the results obtained, and foreshadow shrinkage results in cases where the covariance can be consistently estimated. Suppose then that for a known positive definite matrix Σ we observe X with distribution $\mathcal{N}_d(\theta, \Sigma)$, a normal distribution in \mathbb{R}^d with unknown mean $\theta \in \mathbb{R}^d$. With $f \in W^{1,2}(\nu)$, here taking ν to be the measure of this multivariate Gaussian, we want to compute an unbiased estimate of the mean squared error, or risk, of an estimator of θ of the form $S(x) = x + f(x)$, that is, an unbiased estimate of the expectation of

$$(54) \quad \begin{aligned} & \|S(X) - \theta\|^2 \\ &= \|X - \theta + f(X)\|^2 = \|X - \theta\|^2 + \|f(X)\|^2 + 2\langle f(X), X - \theta \rangle. \end{aligned}$$

Unbiased estimates of the first two terms are easily constructed, as the expectation of the first term is $\text{Tr}(\Sigma)$, a quantity assumed known, and $\|f(X)\|^2$ is an unbiased estimator of its own expectation. Applying the Stein identity (3) to the last term of (54) eliminates the unknown θ via

$$E[\langle X - \theta, f(X) \rangle] = E[\langle \Sigma, \nabla f(X) \rangle]$$

and leads to the conclusion that

$$(55) \quad \text{SURE}(f, X) := \text{Tr}(\Sigma) + \|f(X)\|^2 + 2 \sum_{i,j=1}^d \sigma_{ij} \partial_j f_i(X)$$

is unbiased for the risk; the resulting expression is also computable from the data using the known form of the estimator.

We turn now to the case where X continues to have unknown mean θ and known covariance Σ , but is not necessarily Gaussian. When a Stein kernel $T_{X-\theta}$ exists for X , by applying identity (9) to the final term on the right-hand side of (54) we arrive at the form

$$(56) \quad \text{SURE}_k(f, X) = \text{Tr}(\Sigma) + \|f(X)\|^2 + 2\langle T_{X-\theta}, \nabla f(X) \rangle,$$

which is unbiased for the risk; the subscript k denotes that this version is the Stein kernel form. Alternatively, when the appropriate zero bias vectors exist, from identity (16) we have

$$(57) \quad \text{SURE}_z(f, X) := \text{Tr}(\Sigma) + \|f(X)\|^2 + 2 \sum_{i,j=1}^d \sigma_{ij} \partial_j f_i(X^{ij})$$

is again unbiased for the risk of $S(X)$.

To use the forms (56) and (57) in practice, we would need to be able to generate the Stein kernel, and zero bias vectors, respectively, upon observing X , which is not possible without knowledge of the mean being estimated. Nevertheless, when X is close to normal then heuristically the Stein kernel $T_{X-\theta}$ is close to Σ , and the zero bias vectors X^{ij} , $i, j = 1, \dots, d$ are close in distribution to X . These observations motivate the use of SURE as in (55) with the observed X , which in the approximate normal case should give a risk estimator that is Approximately the Same as SURE (ASSURE), in that it has small bias for the estimate of risk. Propositions 4.1 and 4.2 bound the bias

$$(58) \quad \text{Bias}_\theta(\text{SURE}(f, X)) = E_\theta[\text{SURE}(f, X)] - E_\theta\|S(X) - \theta\|^2$$

of ASSURE, that is, of Stein's unbiased risk estimate (55) when applied in non-Gaussian frameworks.

Proposition 4.1 applies the multivariate Stein kernel framework to determine a bound on the bias of SURE; as in Theorem 3.2, a Stein discrepancy makes an appearance in the bound. A proof is given in Supplement C.

PROPOSITION 4.1. *Let \mathbf{X} have mean $\boldsymbol{\theta}$ and covariance Σ , let $T_{\mathbf{X}-\boldsymbol{\theta}}$ be a Stein kernel for $\mathbf{X} - \boldsymbol{\theta}$ in the sense of (9) and suppose that $\mathbf{f} \in W^{1,2}(\nu)$. Then*

$$(59) \quad |\text{Bias}_{\boldsymbol{\theta}}(\text{SURE}(\mathbf{f}, \mathbf{X}))| \leq 2|E[(\Sigma - T_{\mathbf{X}-\boldsymbol{\theta}}, \nabla \mathbf{f}(\mathbf{X}))]|.$$

If for all $i, j = 1, \dots, d$ the supremum norms $\|\partial_j f_i\|$ over the support of \mathbf{X} are bounded, then letting T_{ij} denote the i, j th entry of $T_{\mathbf{X}-\boldsymbol{\theta}}$,

$$|\text{Bias}_{\boldsymbol{\theta}}(\text{SURE}(\mathbf{f}, \mathbf{X}))| \leq 2 \sum_{i,j=1}^d \|\partial_j f_i\| E[|\sigma_{ij} - T_{ij}|].$$

PROOF. Taking the difference of (55) and (56) yields (59). The second assertion now follows from the first by expanding out the inner product and applying the given bound on the partial derivatives. \square

Proposition 4.1 has the following analog through the use of zero biasing. For $g: \mathbb{R}^d \rightarrow \mathbb{R}$, let $\|g\|_{\text{Lip}}$ denote the usual Lipschitz seminorm of g , and for $i = 1, \dots, d$ let $\|g\|_{\text{Lip},i}$ be the smallest L such that for all real $x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_d$ and u, v ,

$$|g(x_1, \dots, x_{i-1}, u, x_{i+1}, \dots, x_d) - g(x_1, \dots, x_{i-1}, v, x_{i+1}, \dots, x_d)| \leq L|v - u|.$$

PROPOSITION 4.2. *Let $\mathbf{X} = \boldsymbol{\theta} + \mathbf{Y}$ where \mathbf{Y} has mean zero, covariance Σ , and whose zero bias vectors exist. Then, when $\mathbf{f} \in W^{1,2}(\nu)$,*

$$(60) \quad |\text{Bias}_{\boldsymbol{\theta}}(\text{SURE}(\mathbf{f}, \mathbf{X}))| \leq 2 \left| \sum_{i,j=1}^d \sigma_{ij} E_{\boldsymbol{\theta}}(\partial_j f_i(\mathbf{X}^{ij}) - \partial_j f_i(\mathbf{X})) \right|,$$

and when $\partial_j f_i$ is Lipschitz for all $i, j = 1, \dots, d$,

$$(61) \quad |\text{Bias}_{\boldsymbol{\theta}}(\text{SURE}(\mathbf{f}, \mathbf{X}))| \leq 2 \sum_{i,j=1}^d |\sigma_{ij}| \|\partial_j f_i\|_{\text{Lip}} d(\mathbf{X}, \mathbf{X}^{ij}),$$

where $d(\cdot, \cdot)$ is the Wasserstein-1 distance. In addition, under Model 2.3 where $\mathbf{Y}_s, s \in \mathcal{S}$ has independent components and $\Sigma_s = \text{diag}(\sigma_1^2, \dots, \sigma_d^2)$ is nonsingular,

$$(62) \quad |\text{Bias}_{\boldsymbol{\theta}}(\text{SURE}(\mathbf{f}, \mathbf{X}))| \leq 2 \sum_{i=1}^d \sigma_i^2 \|\partial_i f_i\|_{\text{Lip},i} \int_{\mathcal{S}} d(Y_{s,i}^*, Y_{s,i}) d\mu,$$

where $Y_{s,i}^$ has the $Y_{s,i}$ -zero bias distribution.*

Proposition 4.2 is proved in Supplement C. The bounds (59) and (60) above measure deviation from normal through the deviation of the Stein kernel, and zero bias distribution, respectively. Indeed, if the data are normally distributed, then both results return a bound of zero, recovering the Gaussian case.

4.1. *Sure applied to shrinkage.* Now specializing to the shrinkage estimator given by (1), Corollary 4.1 gives two results on the bias of SURE for the shrinkage estimator when applied in non-Gaussian settings, one using Stein kernels, and the other zero biasing. Both claims follow as immediate consequences of our results in Section 3 upon noting that $|\text{Bias}_\theta(\text{SURE}(f, X))|$, as given in (59) and (60), correspond to quantities whose bounds are provided by Theorems 3.2 and 3.4.

COROLLARY 4.1. *Let f be given by (23), $\text{SURE}(f, X)$ as in (55), $\text{Bias}_\theta(\text{SURE}(f, X))$ as in (58) and $X = \theta + Y \in \mathbb{R}^d$, where Y has mean zero and positive definite covariance matrix Σ .*

1. *If Y has Stein kernel T , then with B_λ as given in (32),*

$$|\text{Bias}_\theta(\text{SURE}(f, X))| \leq 2B_\lambda,$$

and if the conditions in (33) hold and $\lambda \in [0, 2(\text{Tr}(\Sigma) - 2\kappa)]$ then this bound is of order $o(d)$.

2. *If the zero bias vectors of Y exist, then with B_λ^* as given in (43),*

$$|\text{Bias}_\theta(\text{SURE}(f, X))| \leq 2B_\lambda^*.$$

The first claim is immediate via Theorem 3.2 and comparing (31) there to (59), and the second claim likewise follows from Theorem 3.4; conditions that guarantee $B_\lambda^* = o(d)$ are detailed following the statement of that latter result.

4.2. *SURE applied to soft-thresholding.* Thresholding of statistical quantities, meaning keeping only “important” quantities, as indicated by their estimates having exceeded some threshold, is widely used in practice. Such procedures are at the core of breakthroughs using wavelet estimation, and their adaptivity in Besov spaces; see [16].

To obtain optimality from the minimax viewpoint, a careful, data-driven selection of the threshold $\lambda > 0$ when estimating the mean of a random vector $X = Y + \theta$, where Y is a centered random vector with covariance matrix $\sigma^2 \text{Id}$, by the estimate $S_\lambda(X)$ whose coordinates are given by soft-thresholding the coordinates of X via $(S_\lambda(x))_i = \text{sgn}(x_i)(|x_i| - \lambda)_+$ for $x_i \in \mathbb{R}$, $\lambda > 0$. By letting $f_\lambda(x) = S_\lambda(x) - x$, the SURE estimate of the risk has the simple following formula:

Outside the Gaussian setting, it is also known that an optimal theoretical value of the threshold gives minimax rates of estimation under some moment assumptions on the noise, in the case of independent coordinates [2, 3, 14]. But in this general framework, to our knowledge, the validity of threshold selection via minimizing the SURE estimate of the associated risks remains an open question.

To begin to approach the problem of obtaining adaptivity results for wavelet estimation using SURE outside the Gaussian setting, we present some conclusions for the selection of a threshold $\lambda > 0$ when estimating the mean of a random vector $X = Y + \theta$, where Y is a centered random vector with covariance matrix $\sigma^2 \text{Id}$, by the estimate $S_\lambda(X)$ whose coordinates are given by soft-thresholding the coordinates of X via $(S_\lambda(x))_i = \text{sgn}(x_i)(|x_i| - \lambda)_+$ for $x_i \in \mathbb{R}$, $\lambda > 0$. By letting $f_\lambda(x) = S_\lambda(x) - x$, the SURE estimate of the risk has the simple following formula:

$$(63) \quad \text{SURE}(f_\lambda, X) = d\sigma^2 + \sum_{i=1}^d \min\{X_i^2, \lambda^2\} - 2\sigma^2 \cdot \text{Card}\{i : |X_i| \leq \lambda\}.$$

Assume now that Y admits a Stein kernel T . By Proposition 4.1, when $f_\lambda \in W^{1,2}(\nu)$,

$$(64) \quad \begin{aligned} |\text{Bias}_\theta(\text{SURE}(f_\lambda, X))| &\leq 2|E_\theta[(\sigma^2 \text{Id} - T, \nabla f_\lambda(X))]| \\ &= 2 \left| \sum_{i=1}^d E_\theta[(\sigma^2 - T_{ii})\mathbf{1}_{\{|X_i| \leq \lambda\}}] \right| \\ &= 2 \left| \sum_{i=1}^d E_\theta[(\sigma^2 - T_{ii})\mathbf{1}_{\{|X_i| > \lambda\}}] \right|, \end{aligned}$$

where the first equality follows from the identities $\partial_i f_{\lambda,j}(X) = -\delta_{ij} \mathbf{1}_{\{|X_i| \leq \lambda\}}$ and the second equality by using the fact that $E[\sigma^2 - T_{ii}] = 0$.

Writing out the expression for the risk of S_λ yields

$$E_\theta[\|S_\lambda(X) - \theta\|^2] = \sum_{i=1}^d E_\theta[(\operatorname{sgn}(X_i)(|X_i| - \lambda)_+ - \theta_i)^2].$$

Reducing to one dimension, for $\theta \in \mathbb{R}$, letting

$$p(\lambda, \theta) = E[(\operatorname{sgn}(Y + \theta)(|Y + \theta| - \lambda)_+ - \theta)^2],$$

one may verify that

$$(65) \quad p(\lambda, \theta) \geq P(|Y + \theta| > \lambda + |\theta| + 1).$$

The latter bound will be useful for controlling the bias of SURE by the risk in the strongly log-concave case.

4.2.1. Strongly log-concave case. Assume furthermore that Y has a positive strongly log-concave density on \mathbb{R}^d . In this case, T is uniformly bounded [20], so there exists a positive constant L such that $\max_i |\sigma^2 - T_{ii}| \leq L$ a.s. In addition, Y has sub-Gaussian tails, in the sense that there exists a constant $a > 0$ such that for any $t > 0$, $P(\|Y\| \geq t) \leq ae^{-t^2/C}$. Note that when such property is in force, it is sufficient according to [2], to search among values of λ in the range $I = [0, \sqrt{C \log d}]$. Hence, inequalities (64) and (65) give

$$(66) \quad |\operatorname{Bias}_\theta(\operatorname{SURE}(f_\lambda, X))| \leq 2L \sum_{i=1}^d P(|X_i| > \lambda) \leq 2LM \cdot E_\theta[\|S_\lambda(X) - \theta\|^2],$$

$$\text{where } M = \sup_{\lambda \in I, i=1, \dots, d} \frac{P(|X_i| > \lambda)}{P(|X_i| > \lambda + A + 1)} \text{ and } A = \|\theta\|_\infty.$$

We are now ready to state our main result about adaptive soft-thresholding calibration. Theorem 4.1 is proved in Supplement C.

THEOREM 4.1. Assume that $X = \theta + Y$ with Y mean zero and covariance matrix $\sigma^2 \operatorname{Id}$, and has a strongly log-concave distribution with independent coordinates, and that $LM < 1/2$, with the constants L and M defined above. Consider the selection of the soft-thresholding parameter via SURE,

$$(67) \quad \hat{\lambda} \in \arg \min_{\lambda \in I} \operatorname{SURE}(f_\lambda, X),$$

where $I = [0, \sqrt{C \log d}]$, for C the scaling sub-Gaussian constant of Y . Then

$$(1 - 2LM) E_\theta[\|\hat{S}_\lambda(X) - \theta\|^2] \leq (1 + 2LM) \min_{\lambda \in I} E_\theta[\|S_\lambda(X) - \theta\|^2] + B\sqrt{d \log^3 d},$$

where B is a positive constant depending only on C .

In the setting of wavelet estimation, one considers a vector of coefficients computed by applying the wavelet transform on an input signal vector. But independence of the noise terms in the signal is lost in general—except in the Gaussian case—when taking linear combinations of the transformed signal. This difficulty would be the first one to overcome in order to generalize our results to the wavelet estimation problem.

4.3. *Adaptivity under classical asymptotics.* Let us consider a d -dimensional vector X with mean θ and variance $\sigma_d^2 \text{Id}$ where $\sigma_d^2 = \sigma^2/d$, with σ^2 an absolute positive constant. This is a case of interest in statistics, where the vector X might be the mean of d i.i.d. vectors with finite variance σ^2 . This setting is also naturally linked to a nonparametric regression model; see, for instance, [45], Section 7.3.

Pinsker's theorem [38] (see also [36]) gives the exact asymptotic minimax risk over ℓ_2 -balls in the Gaussian case. More precisely, letting $\mathcal{G}_d(c) = \{\mathcal{N}(\theta, \sigma_d^2 \text{Id}) : \|\theta\| \leq c\}$, we have

$$(68) \quad \lim_{d \rightarrow +\infty} \inf_{\hat{\theta}} \sup_{P \in \mathcal{G}_d(c)} E_P[\|\hat{\theta} - \theta\|^2] = \frac{\sigma^2 c^2}{\sigma^2 + c^2},$$

where the infimum is taken over all estimators of θ , that is, over all measurable functions of X for which $\theta = E_P[X]$.

The asymptotic value of the Gaussian minimax risk can actually be extended to the whole class of distributions $\mathcal{P}_d(c) = \{P \in \mathcal{M}_1^+ : \|E_P[X]\| \leq c, \text{Var}_P[X] = \sigma_d^2 \text{Id}\}$, where \mathcal{M}_1^+ is the set of all probability measures on \mathbb{R}^d . More precisely, for any collection of distributions \mathcal{P} such that $\mathcal{G}_d(c) \subset \mathcal{P} \subset \mathcal{P}_d(c)$, it holds

$$(69) \quad \lim_{d \rightarrow +\infty} \inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} E_P[\|\hat{\theta} - \theta\|^2] = \frac{\sigma^2 c^2}{\sigma^2 + c^2}.$$

Indeed, by (68) the left-hand side of (69) is at least as large as the right, since $\mathcal{G}_d(c) \subset \mathcal{P}$. The reverse inequality is achieved by considering the estimator $\hat{\theta} = c^2 X / (\sigma^2 + c^2)$, which satisfies $E_P[\|\hat{\theta} - \theta\|^2] \leq \sigma^2 c^2 / (\sigma^2 + c^2)$ whenever $P \in \mathcal{P}_d(c)$.

In the Gaussian case, the James–Stein estimator $S_\lambda(X)$ in (1), with $\lambda = (d - 2)\sigma_d^2$, is known to be adaptive, in the sense that it asymptotically recovers the minimax risk for any $c > 0$, without the knowledge of c . Hence, a natural question is: under what more general distributional assumptions is the James–Stein estimator adaptive to c ? That is, for which collections of distributions $\{\mathcal{P}_c : c > 0\}$, where $\mathcal{G}_d(c) \subset \mathcal{P}_c \subset \mathcal{P}_d(c)$, does the James–Stein estimator recover the asymptotic minimax risk for any fixed value of $c > 0$? We answer this question with the following results, starting with the use of Stein kernels.

We note that a variance decay of rate σ^2/d corresponds to a decay on the error of the form Y/\sqrt{d} , and in the proof of the following result, that is given in Supplement C, we invoke Theorem 3.2 with that form for the error. Note, correspondingly, that scaling corresponds to a decay on the Stein kernel T of Y to T/d .

THEOREM 4.2. *Let $X - \theta$ be a mean zero vector with covariance matrix $\sigma_d^2 \text{Id}$ with $\sigma_d^2 = \sigma^2/d$, and let T/d be a Stein kernel for $X - \theta$ in the sense of (9). Then, with $\lambda = (d - 2)\sigma_d^2$, it holds that*

$$(70) \quad E[\|S_\lambda(X) - \theta\|^2] \leq d\sigma_d^2 - \frac{(d - 2)^2 \sigma_d^4}{\|\theta\|^2 + d\sigma_d^2} + 2B_\lambda,$$

where B_λ is as in (32). If the conditions in (33) or in (40) hold for X and T , then B_λ is of order $o(1)$.

Again Lemma 6.1 can be applied to obtain bounds on the expectations of the inverse moments under Model 2.3. In the case of a log-concave vector, Theorem 4.2 gives the following corollary, proved in Supplement C.

COROLLARY 4.2. *Let $\mathcal{P}(c)$ be the set of distributions of vectors X belonging to the set $\mathcal{P}_d(c)$ defined above, such that $X - \theta$ is a mean zero isotropic log-concave vector. Then the James–Stein estimator $S_\lambda(X)$ in (1), with $\lambda = (d - 2)\sigma_d^2$, is asymptotically adaptive to c in the set $\mathcal{P}(c)$, in the sense that it recovers the minimax risk over $\mathcal{P}(c)$ for any $c > 0$, without the knowledge of c , when the dimension d grows to infinity.*

We now turn to using the zero bias distribution to obtain parallel results. For $\boldsymbol{\theta} \in \mathbb{R}^d$ and C_4, C_{-2} positive constants, let $\mathcal{P}(\boldsymbol{\theta}, C_4, C_{-2})$ be the set of distributions of vectors $\mathbf{X} = \mathbf{Y} + \boldsymbol{\theta}$ that satisfy the assumptions of Part 2 of Theorem 3.4, where $\mathbf{Y}_s, s \in \mathcal{S}$ has covariance matrix $\sigma_d^2 \text{Id}$ and $\sup_{s \in \mathcal{S}} \max_{1 \leq i \leq d} E[Y_{s,i}^4] \leq C_4$.

THEOREM 4.3. *If the distribution of \mathbf{X} is a member of $\mathcal{P}(\boldsymbol{\theta}, C_4, C_{-2})$, then with $\lambda = (d - 2)\sigma_d^2$,*

$$(71) \quad E[\|S_\lambda(\mathbf{X}) - \boldsymbol{\theta}\|^2] \leq \frac{\sigma^2 \|\boldsymbol{\theta}\|^2}{\sigma^2 + \|\boldsymbol{\theta}\|^2} \left(1 + \frac{4\sigma^2}{d\|\boldsymbol{\theta}\|^2}\right) + L\lambda \left(\frac{1}{d} + \frac{\|\boldsymbol{\theta}\|_1}{d^2} + \frac{\|\boldsymbol{\theta}\|_2^2}{d^3}\right),$$

where the constant L only depends on σ^2, C_4 and C_{-2} . Moreover, letting

$$\mathcal{P}(c) = \{P \in \mathcal{P}(\boldsymbol{\theta}, C_4, C_{-2}) : \|\boldsymbol{\theta}\| \leq c\}$$

the James–Stein estimator $S_\lambda(\mathbf{X})$ in (1) is asymptotically adaptive to c in the set $\mathcal{P}(c)$, in the sense that it recovers the minimax risk over $\mathcal{P}(c)$ for any $c > 0$, without the knowledge of c , when the dimension d grows to infinity.

Theorem 4.3 shows that the James–Stein estimator is adaptive in this case, in the sense that it asymptotically recovers the minimax risk over ℓ_2 -balls, without requiring that c be known. Its proof can be found in Supplement C.

5. Multivariate zero bias. We collect the properties of the zero bias distribution in the following result, proved in Supplement D. We first note that when \mathbf{Y} satisfies (15) its mean is necessarily zero.

PROPOSITION 5.1. *Let $\mathbf{Y} \in \mathbb{R}^d$ have mean zero and positive definite covariance matrix Σ .*

1. *If (15) holds, then $\Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_d^2)$, and the laws for random vectors $\mathbf{Y}^i, i = 1, \dots, d$ satisfying (14) exist and are unique. Conversely, if (14) holds then (15) holds.*

When (15) holds, the collection of zero bias random vectors may be constructed as follows. With ν the distribution of \mathbf{Y} , for each $i = 1, \dots, d$ let $\mathbf{Y}^{\square,i}$ have distribution

$$(72) \quad d\nu^{\square,i} = \frac{y_i^2}{\sigma_i^2} d\nu.$$

For $\mathbf{y} \in \mathbb{R}^d, u \in \mathbb{R}$ and $i = 1, \dots, d$, let

$$D_{i,u}\mathbf{y} = (y_1, \dots, y_{i-1}, uy_i, y_{i+1}, \dots, y_d)^\top,$$

that is, $D_{i,u}\mathbf{y}$ is formed by multiplying the i th component of \mathbf{y} by u . Then for U_i a uniformly distributed variable on $[0, 1]$, independent of $\mathbf{Y}^{\square,i}$, the collection of vectors

$$(73) \quad \mathbf{Y}^i = D_{i,U_i}\mathbf{Y}^{\square,i} \quad \text{for } i = 1, \dots, d$$

satisfies (14).

2. *When \mathbf{Y} satisfies (15), and S and S^i are the supports of \mathbf{Y} and \mathbf{Y}^i , respectively, then with cl denoting closure,*

$$(74) \quad S^i = \text{cl}(U^i(S)) \quad \text{where } U^i(S) = \{D_{i,u}\mathbf{y} : \mathbf{y} \in S, y_i \neq 0, u \in [0, 1]\}.$$

3. When $\mathbf{Y} = \sum_{j=1}^n \mathbf{Y}_j$ where $\mathbf{Y}_j, j = 1, \dots, n$ are independent, mean zero \mathbb{R}^d valued random vectors with covariance matrices $\Sigma_j = \text{diag}(\sigma_{j,1}^2, \dots, \sigma_{j,d}^2)$ and associated zero bias vectors $\mathbf{Y}_j^i, i = 1, \dots, d$, then \mathbf{Y} has zero bias vectors $\mathbf{Y}^i, i = 1, \dots, d$ whose distributions are the mixtures of $\mathbf{Y} - \mathbf{Y}_j + \mathbf{Y}_j^i, j = 1, \dots, n$, where \mathbf{Y}_j^i is the i th zero bias vector of \mathbf{Y}_j , taken independently of $\mathbf{Y}_k, k \neq j$ with probability $\sigma_{j,i}^2 / \sigma_i^2$, where $\sigma_i^2 = \sum_{j=1}^n \sigma_{j,i}^2$.

4. When \mathbf{Y} is the μ mixture of $\{\mathbf{Y}_s\}_{s \in \mathcal{S}}$, a collection of mean zero random vectors in \mathbb{R}^d with $\mathbf{Y}_s, s \in \mathcal{S}$ having nonsingular covariance matrices $\text{Var}(\mathbf{Y}_s) = \text{diag}(\sigma_{s,1}^2, \dots, \sigma_{s,d}^2)$ and zero bias vectors $\mathbf{Y}_s^i, i = 1, \dots, d$, then the zero bias distribution of \mathbf{Y} exists, and the distribution of \mathbf{Y}^i is the v^i mixture of $\mathbf{Y}_s^i, s \in \mathcal{S}$, where $dv^i / d\mu = \sigma_{s,i}^2 / \sigma_i^2$ where $\sigma_i^2 = \text{Var}(Y_i)$. In particular, $v^i = \mu$ if and only if $\sigma_{s,i}^2$ is a constant μ a.s. over $s \in \mathcal{S}$.

5. When $\mathbf{Y} = \mathbf{A}\mathbf{U} \in \mathbb{R}^d$ for some mean zero $\mathbf{U} \in \mathbb{R}^m$ with positive definite covariance matrix Γ and whose zero bias vectors \mathbf{U}^{kl} exist for all $1 \leq k, l \leq m$ for which $\gamma_{kl} \neq 0$, $\mathbf{A} = (a_{ik})_{1 \leq i \leq d, 1 \leq k \leq m} \in \mathbb{R}^{d \times m}$ and $\sigma_{ij} := \text{Cov}(Y_i, Y_j) \geq 0$ for all $1 \leq i, j \leq d$, and for all i, j such that $\sigma_{ij} > 0$ we have $a_{ik}\gamma_{kl}a_{jl} > 0$ for $1 \leq k, l \leq m$, then the zero bias vectors for \mathbf{Y} exist, with the distribution of \mathbf{Y}^{ij} for such i, j pairs obtained by mixing the distributions of $\mathbf{A}\mathbf{U}^{kl}$ with measure $\mu_{ij}(kl) = a_{ik}\gamma_{kl}a_{jl} / \sigma_{ij}$.

6. When \mathbf{Y} satisfies (15) and has density $p(\mathbf{y})$ then for all $i = 1, \dots, d$, the integral

$$(75) \quad p^i(\mathbf{y}) = \frac{1}{\sigma_i^2} \int_{y_i}^{\infty} up(y_1, \dots, y_{i-1}, u, y_{i+1}, \dots, y_d) du$$

exists a.e. and $p^i(\cdot)$ is the density of \mathbf{Y}^i . If there exists $g \in L^1(\mathbb{R})$ such that $|y_i|p(\mathbf{y}) \leq g(y_i)$, then $p^i(\mathbf{y})$ is bounded over \mathbb{R}^d .

Taking $\mathbf{U} \sim \mathcal{U}[0, 1]$ and \mathbf{Y} independent, item 1 provides the alternative identity

$$E[f(\mathbf{Y}^i)] = \frac{1}{\sigma_i^2} E[Y_i^2 f(D_{i,U} \mathbf{Y})]$$

to (16) for computing expectations with respect to \mathbf{Y}^i . Generally, when \mathbf{X}^i exists for $\mathbf{X} = \mathbf{Y} + \boldsymbol{\theta}$, we obtain

$$(76) \quad E[f(\mathbf{X}^i)] = \frac{1}{\sigma_i^2} E[(X_i - \theta_i)^2 f(D_{i,U}(\mathbf{X} - \boldsymbol{\theta}) + \boldsymbol{\theta})].$$

The necessity for excluding $y_i = 0$ in (74) can be made apparent by considering the zero bias distribution of the measure in \mathbb{R}^2 that puts equal mass on the five points $(\pm 1, \pm 1)$ and $(0, 0)$, and for the necessity of taking the closure, consider \mathbf{Y} with the uniform distribution on the boundary of the L^1 ball in \mathbb{R}^2 .

6. Boundedness in mean of inverse norms. In this section, we present two results to bound the expectation of powers of inverse norms of a vector \mathbf{X} , a quantity on which our results here depend. The first provides a simple sufficient condition on the moment generating function of the squared norm of the vector whose inverse moment is being taken, and the other that can be applied when the vector has a log-concave distribution. The proof of Lemma 6.1 can be found in Supplement E.

Before stating the first result, we recall that random variables V_1, \dots, V_d are said to be negatively associated (see [27]) when for all disjoint subsets A and B of $\{1, \dots, d\}$,

$$\text{Cov}(f(V_i, i \in A), g(V_i, i \in B)) \leq 0$$

when f and g are both nondecreasing (or both nonincreasing) functions. Clearly, collections of independent random variables are negatively associated.

LEMMA 6.1. Let $S_d, d \geq 1$ be a nonnegative random variable such that for some μ, q and C , all positive,

$$(77) \quad M_d(t) \leq \frac{C}{(1 - \mu t/q)^{qd}} \quad \text{for all } t \leq 0, \text{ where } M_d(t) = E[e^{tS_d}].$$

Then for all $m \geq 1$, if $d \geq 2m/q$ there exists a constant $C_{\mu,m}$, depending only on μ and m such that

$$E\left[\frac{d}{S_d}\right]^m \leq C_{\mu,m}.$$

When $S_d = \sum_{i=1}^d V_i$, a sum of nonnegative, negatively associated random variables such that for some μ and q positive the moment generating functions of $V_i, i = 1, \dots, d$ obey the bound (77) with $C = 1$ and $d = 1$, then (35)a holds for all $d \geq 2m/q$.

REMARK 6.1. When $S_d = \|X\|^2$ for $X \in \mathbb{R}^d$, then Lemma 6.1 provides a sufficient condition for the satisfaction of the negative moment condition (35)a in terms of the moment generating function of S_d . The lemma can be applied to vectors having negatively associated components, and hence in particular to those with independent components, and then by the extension as done in Corollary 3.1, for vectors having nonindependent coordinate distributions that are covered by Model 2.3.

For instance, when the marginal distribution $\mathcal{L}(Y)$ of the components of the error vector Y is $\mathcal{N}(0, \sigma^2)$, then $V = (\theta + Y)^2$ has a noncentral χ^2 distribution with moment generating function

$$M_V(t) = E[e^{t(\theta+Y)^2}] = \frac{\exp(\frac{t\theta^2}{1-2t\sigma^2})}{(1-2t\sigma^2)^{1/2}}.$$

As the numerator is bounded by 1 for all $t \leq 0$, and $M_V(t)$ does not otherwise depend on θ , we see that (77) is satisfied for all $\theta \in \mathbb{R}$ with $C = 1$, $q = 1/2$, $\mu = \sigma^2$ and $d = 1$. If the distribution of the absolute value of θ plus the coordinate error Y stochastically dominates the same quantity for a σ scaling of the standard normal, that is, when $|\theta + \sigma Z| \leq_{\text{st}} |\theta + Y|$, the same bound will hold.

Though condition (77) appears related to the subgamma property of random variables, that condition is concerned about the behavior of the moment generating function in a positive neighborhood of zero. Note also that when any mass of a distribution is moved to zero it only would “help” the satisfaction of the subgamma property, but (77) will immediately be violated. Indeed, if V has a point mass of probability $p > 0$ at zero, then $M_V(t) \geq p$ for all $t \leq 0$, and hence $M_V(t)$ cannot tend to zero as $t \rightarrow -\infty$, as does any moment generating function that satisfies (77).

REMARK 6.2. We continue the discussion in Example 2.3 where the error vector Y has a discrete distribution with finite support S_Y . Taking any $\delta \in (0, 1]$ and

$$(78) \quad \theta \in \bigcap_{y \in S_Y} \bigcup_{I \subset \{1, \dots, d\}, |I| \geq \delta d} \{\psi : \psi_i \neq -y_i, i \in I\},$$

for any $y \in S_Y$ there must exist a set of indices $I \subset \{1, \dots, d\}$ satisfying $|I| \geq \delta d$ such that

$$\tau_y := \min_{i \in I} |\theta_i + y_i| > 0.$$

Now, letting $\tau = \min_{y \in S_Y} \tau_y$, a quantity that must be positive as S_Y is finite, we obtain $\|X\|^2 = \|\theta + Y\|^2 \geq \delta d \tau^2$ almost surely. With only minor changes to the argument to handle X^{-i} , we see that the negative moment condition (35)b is satisfied. Similar to the conclusion in Example 2.3, the exceptional set of shifts θ not satisfying (78) has Lebesgue measure zero.

REMARK 6.3. The assumption that there exists a positive constant K such that for any $a \geq 0$ and any $i \in \{1, \dots, d\}$,

$$\sup_{u \in \mathbb{R}} P\{u - a \leq Y_i \leq u + a\} \leq Ka$$

is made in the proof of Theorem 3.1 in [32] in order to tackle negative moment estimates, in the case where the components of the observation vector are independent, an instance subsumed by Model 2.3. The main example achieving this condition is the case where $Y_i, i = 1, \dots, d$ has a uniformly bounded density with respect to Lebesgue measure, with a bound independent of i . In such a case, the vector Y also has a bounded density on \mathbb{R}^d , since its coordinates are independent, and negative moments $E[\|X\|^{-2m}]$ are finite for any $m \geq 1$ for dimensions $d \geq 2m + 1$. We note that [32] does not need to control a decay rate of these moments with respect to the ambient dimension—as we do for instance in (35)—but there are two essential differences between our analysis and that in [32]: first, the setting of [32] is asymptotic only, and second, [32] considers the consistency in probability of SURE toward the loss, while we investigate the bias of SURE compared to the risk, which is the integrated loss.

We now give a bound on means of inverse norms for log-concave random vectors, which is a corollary of [1], Theorem 6.2 (itself a variant of results of [37]).

PROPOSITION 6.1. *There is a dimension d_0 such that for any log-concave distribution in \mathbb{R}^d with covariance matrix $\sigma^2 \text{Id}$ for some $\sigma^2 > 0$, for $d \geq d_0$ we have*

$$E[\|X\|^{-6}] \leq cd^{-3},$$

where c is a constant independent of the distribution and of d .

In this statement, the exponent 6 does not play a significant role, beyond affecting the value of c and the dimension d_0 . The value of d_0 depends on the values of some universal constants used in [1], that were not made explicit, though it must be larger than 6.

Strictly speaking, [1], Theorem 6.2, is only given for centered random variables. In the noncentered case, we can consider the projection \tilde{X} of X onto the $(d - 1)$ -dimensional subspace orthogonal to the mean vector θ . Then \tilde{X} is a $(d - 1)$ -dimensional centered log-concave vector, still with covariance matrix $\sigma^2 \text{Id}$. We can then apply the centered result to \tilde{X} , and use the fact that $\|X\|^{-8} \leq \|\tilde{X}\|^{-8}$.

Acknowledgments. This research was started during the workshop *Stein’s method and applications in high-dimensional statistics* at the American Institute of Mathematics (AIM), San Jose, California, and we acknowledge the generous support and the fertile research environment provided by AIM. The authors sincerely thank the Associate Editor and our two reviewers for their hard work in providing us with detailed and insightful reviews. The fourth author warmly thanks Lionel Truquet for instructive discussions related to martingale difference random fields.

Funding. MF was additionally supported by ANR-11-LABX-0040-CIMI within the program ANR-11-IDEX-0002-02, as well as Projects EFI (ANR-17-CE40-0030) and MESA (ANR-18-CE40-006) of the French National Research Agency (ANR).

GR is partially supported by EPSRC grants EP/R018472/1 and EP/T018445/1.

SUPPLEMENTARY MATERIAL

Supplement to “Relaxing the Gaussian assumption in shrinkage and SURE in high dimension” (DOI: [10.1214/22-AOS2208SUPP](https://doi.org/10.1214/22-AOS2208SUPP); .pdf). The supplement contains proofs of the main results and technical details for some remarks and examples. Sections A to E of the supplement correspond to Sections 2 to 6 of the paper, and Section F provides insights towards generalizing our results to situations other than shrinkage via extensions of Assumption 3.1.

REFERENCES

- [1] ADAMCZAK, R., GUÉDON, O., LATAŁA, R., LITVAK, A. E., OLESZKIEWICZ, K., PAJOR, A. and TOMCZAK-JAEGERMANN, N. (2012). Moment estimates for convex measures. *Electron. J. Probab.* **17** no. 101, 19. [MR3005719](https://doi.org/10.1214/EJP.v17-2150) <https://doi.org/10.1214/EJP.v17-2150>
- [2] AVERKAMP, R. and HOUDRÉ, C. (2003). Wavelet thresholding for non-necessarily Gaussian noise: Idealism. *Ann. Statist.* **31** 110–151. [MR1962501](https://doi.org/10.1214/aos/1046294459) <https://doi.org/10.1214/aos/1046294459>
- [3] AVERKAMP, R. and HOUDRÉ, C. (2005). Wavelet thresholding for nonnecessarily Gaussian noise: Functionality. *Ann. Statist.* **33** 2164–2193. [MR2211083](https://doi.org/10.1214/009053605000000471) <https://doi.org/10.1214/009053605000000471>
- [4] BAYATI, M., ERDOĞDU, M. and MONTANARI, A. (2013). Estimating lasso risk and noise level. In *Advances in Neural Information Processing Systems*.
- [5] BOBKOV, S. G. and NAZAROV, F. L. (2003). On convex bodies and log-concave probability measures with unconditional basis. In *Geometric Aspects of Functional Analysis. Lecture Notes in Math.* **1807** 53–69. Springer, Berlin. [MR2083388](https://doi.org/10.1007/978-3-540-36428-3_6) https://doi.org/10.1007/978-3-540-36428-3_6
- [6] CACOULOS, T. and PAPATHANASIOU, V. (1992). Lower variance bounds and a new proof of the central limit theorem. *J. Multivariate Anal.* **43** 173–184. [MR1193610](https://doi.org/10.1016/0047-259X(92)90032-B) [https://doi.org/10.1016/0047-259X\(92\)90032-B](https://doi.org/10.1016/0047-259X(92)90032-B)
- [7] CANDÈS, E. J., SING-LONG, C. A. and TRZASKO, J. D. (2013). Unbiased risk estimates for singular value thresholding and spectral estimators. *IEEE Trans. Signal Process.* **61** 4643–4657. [MR3105401](https://doi.org/10.1109/TSP.2013.2270464) <https://doi.org/10.1109/TSP.2013.2270464>
- [8] CATTIAUX, P. and GUILLIN, A. (2020). On the Poincaré constant of log-concave measures. In *Geometric Aspects of Functional Analysis. Vol. I. Lecture Notes in Math.* **2256** 171–217. Springer, Cham. [MR4175748](https://doi.org/10.1007/978-3-030-36020-7_9) https://doi.org/10.1007/978-3-030-36020-7_9
- [9] CELLIER, D., FOURDRINIER, D. and ROBERT, C. (1989). Robust shrinkage estimators of the location parameter for elliptically symmetric distributions. *J. Multivariate Anal.* **29** 39–52. [MR0991055](https://doi.org/10.1016/0047-259X(89)90075-4) [https://doi.org/10.1016/0047-259X\(89\)90075-4](https://doi.org/10.1016/0047-259X(89)90075-4)
- [10] CHATTERJEE, S. (2009). Fluctuations of eigenvalues and second order Poincaré inequalities. *Probab. Theory Related Fields* **143** 1–40. [MR2449121](https://doi.org/10.1007/s00440-007-0118-6) <https://doi.org/10.1007/s00440-007-0118-6>
- [11] CHEN, Y. (2021). An almost constant lower bound of the isoperimetric coefficient in the KLS conjecture. *Geom. Funct. Anal.* **31** 34–61. [MR4244847](https://doi.org/10.1007/s00039-021-00558-4) <https://doi.org/10.1007/s00039-021-00558-4>
- [12] CHEN, Y., WIESEL, A. and HERO, A. O. III (2011). Robust shrinkage estimation of high-dimensional covariance matrices. *IEEE Trans. Signal Process.* **59** 4097–4107. [MR2865971](https://doi.org/10.1109/TSP.2011.2138698) <https://doi.org/10.1109/TSP.2011.2138698>
- [13] COURTADE, T. A., FATHI, M. and PANANJADY, A. (2019). Existence of Stein kernels under a spectral gap, and discrepancy bounds. *Ann. Inst. Henri Poincaré Probab. Stat.* **55** 777–790. [MR3949953](https://doi.org/10.1214/18-aihp898) <https://doi.org/10.1214/18-aihp898>
- [14] DELYON, B. and JUDITSKY, A. (1996). On minimax wavelet estimators. *Appl. Comput. Harmon. Anal.* **3** 215–228. [MR1400080](https://doi.org/10.1006/acha.1996.0017) <https://doi.org/10.1006/acha.1996.0017>
- [15] DÖBLER, C. (2015). New Berry–Esseen and Wasserstein bounds in the CLT for non-randomly centered random sums by probabilistic methods. *ALEA Lat. Am. J. Probab. Math. Stat.* **12** 863–902. [MR3453299](https://doi.org/10.1214/15-ALEA0000) <https://doi.org/10.1214/15-ALEA0000>
- [16] DONOHO, D. L. and JOHNSTONE, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.* **90** 1200–1224. [MR1379464](https://doi.org/10.1080/01621459.1995.10476844) <https://doi.org/10.1080/01621459.1995.10476844>
- [17] ELDAR, Y. C. (2009). Generalized SURE for exponential families: Applications to regularization. *IEEE Trans. Signal Process.* **57** 471–481. [MR2603376](https://doi.org/10.1109/TSP.2008.2008212) <https://doi.org/10.1109/TSP.2008.2008212>
- [18] EVANS, L. C. (1998). *Partial Differential Equations. Graduate Studies in Mathematics* **19**. Amer. Math. Soc., Providence, RI. [MR1625845](https://doi.org/10.1214/18-AOP1305) <https://doi.org/10.1214/18-AOP1305>
- [19] EVANS, S. N. and STARK, P. B. (1996). Shrinkage estimators, Skorokhod’s problem and stochastic integration by parts. *Ann. Statist.* **24** 809–815. [MR1394989](https://doi.org/10.1214/aos/1032894466) <https://doi.org/10.1214/aos/1032894466>
- [20] FATHI, M. (2019). Stein kernels and moment maps. *Ann. Probab.* **47** 2172–2185. [MR3980918](https://doi.org/10.1214/18-AOP1305) <https://doi.org/10.1214/18-AOP1305>

- [21] FATHI, M., GOLDSTEIN, L., REINERT, G. and SAUMARD, A. (2022). Supplement to “Relaxing the Gaussian assumption in shrinkage and SURE in high dimension.” <https://doi.org/10.1214/22-AOS2208SUPP>
- [22] FOURDRINIER, D., STRAWDERMAN, W. E. and WELLS, M. T. (2018). *Shrinkage Estimation*. Springer Series in Statistics. Springer, Cham. MR3887633 <https://doi.org/10.1007/978-3-030-02185-6>
- [23] GOLDSTEIN, L. (2007). L^1 bounds in normal approximation. *Ann. Probab.* **35** 1888–1930. MR2349578 <https://doi.org/10.1214/009117906000001123>
- [24] GOLDSTEIN, L. and REINERT, G. (1997). Stein’s method and the zero bias transformation with application to simple random sampling. *Ann. Appl. Probab.* **7** 935–952. MR1484792 <https://doi.org/10.1214/aoap/1043862419>
- [25] GOLDSTEIN, L. and REINERT, G. (2005). Zero biasing in one and higher dimensions, and applications. In *Stein’s Method and Applications. Lect. Notes Ser. Inst. Math. Sci. Natl. Univ. Singap.* **5** 1–18. Singapore Univ. Press, Singapore. MR2201883 https://doi.org/10.1142/9789812567673_0001
- [26] JAMES, W. and STEIN, C. (1961). Estimation with quadratic loss. In *Proc. 4th Berkeley Sympos. Math. Statist. and Prob., Vol. I* 361–379. Univ. California Press, Berkeley, CA. MR0133191
- [27] JOAG-DEV, K. and PROSCHAN, F. (1983). Negative association of random variables, with applications. *Ann. Statist.* **11** 286–295. MR0684886 <https://doi.org/10.1214/aos/1176346079>
- [28] KLARTAG, B. (2013). Poincaré inequalities and moment maps. *Ann. Fac. Sci. Toulouse Math.* (6) **22** 1–41. MR3247770 <https://doi.org/10.5802/afst.1366>
- [29] LANDSMAN, Z., VANDUFFEL, S. and YAO, J. (2015). Some Stein-type inequalities for multivariate elliptical distributions and applications. *Statist. Probab. Lett.* **97** 54–62. MR3299751 <https://doi.org/10.1016/j.spl.2014.11.005>
- [30] LEDOUX, M., NOURDIN, I. and PECCATI, G. (2015). Stein’s method, logarithmic Sobolev and transport inequalities. *Geom. Funct. Anal.* **25** 256–306. MR3320893 <https://doi.org/10.1007/s00039-015-0312-0>
- [31] LEE, Y. T. and VEMPALA, S. (2017). The KLS conjecture. In *Current Developments in Mathematics*.
- [32] LI, K.-C. (1985). From Stein’s unbiased risk estimates to the method of generalized cross validation. *Ann. Statist.* **13** 1352–1377. MR0811497 <https://doi.org/10.1214/aos/1176349742>
- [33] MIJOULE, G., REINERT, G. and SWAN, Y. (2018). Stein operators, kernels and discrepancies for multivariate continuous distributions. [arXiv:1806.03478](https://arxiv.org/abs/1806.03478).
- [34] NAHAPETIAN, B. S. and PETROSIAN, A. N. (1992). Martingale-difference Gibbs random fields and central limit theorem. *Ann. Acad. Sci. Fenn. Ser. A I Math.* **17** 105–110. MR1162153 <https://doi.org/10.5186/aasfm.1992.1713>
- [35] NOURDIN, I. and PECCATI, G. (2012). *Normal Approximations with Malliavin Calculus: From Stein’s Method to Universality*. Cambridge Tracts in Mathematics **192**. Cambridge Univ. Press, Cambridge. MR2962301 <https://doi.org/10.1017/CBO9781139084659>
- [36] NUSSBAUM, M. (1996). The Pinsker bound: A review. In *Encyclopedia of Statistical Sciences* (S. Kotz, ed.) Wiley, New York, NY.
- [37] PAOURIS, G. (2012). Small ball probability estimates for log-concave measures. *Trans. Amer. Math. Soc.* **364** 287–308. MR2833584 <https://doi.org/10.1090/S0002-9947-2011-05411-5>
- [38] PINSKER, M. S. Optimal filtration of square-integrable signals in Gaussian noise. MR0624591
- [39] SAUMARD, A. and WELLNER, J. A. (2014). Log-concavity and strong log-concavity: A review. *Stat. Surv.* **8** 45–114. MR3290441 <https://doi.org/10.1214/14-SS107>
- [40] SRIVASTAVA, M. S. and BILODEAU, M. (1989). Stein estimation under elliptical distributions. *J. Multivariate Anal.* **28** 247–259. MR0991949 [https://doi.org/10.1016/0047-259X\(89\)90108-5](https://doi.org/10.1016/0047-259X(89)90108-5)
- [41] STEIN, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 1954–1955, Vol. I 197–206. Univ. California Press, Berkeley–Los Angeles, CA. MR0084922
- [42] STEIN, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist.* **9** 1135–1151. MR0630098
- [43] UTEV, S. A. (1989). Probabilistic problems connected with an integro-differential inequality. *Sibirsk. Mat. Zh.* **30** 182–186, 220. MR1010851 <https://doi.org/10.1007/BF00971508>
- [44] VEMPALA, S. (2010). Recent progress and open problems in algorithmic convex geometry. In *IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science (FSTTCS 2010)*.
- [45] WASSERMAN, L. (2006). *All of Nonparametric Statistics*. Springer Texts in Statistics. Springer, New York. MR2172729
- [46] ZHANG, X.-P. and MITA, D. (1998). Adaptive denoising based on SURE risk. *IEEE Signal Process. Lett.* **5** 265–267.