Improved Estimation of Dissimilarities by Pre-smoothing Functional Data

David B. Hitchcock
University of South Carolina

George Casella
University of Florida

James G. Booth
Cornell University*

June 9, 2005

Abstract

We examine the effect of pre-smoothing functional data on estimating the dissimilarities among objects in a data set, with applications to cluster analysis and other distance methods such as multidimensional scaling and statistical matching. We prove that a shrinkage

*David B. Hitchcock is Assistant Professor, Department of Statistics, University of South Carolina, Columbia, SC 29208 (email: hitchcock@stat.sc.edu). George Casella is Distinguished Professor and Chair, Department of Statistics, University of Florida, Gainesville, FL 32611 (email: casella@stat.ufl.edu). James Booth is Professor, Biological Statistics and Computational Biology, Cornell University, Ithaca, NY 14853 (email: jb383@cornell.edu). This work was partially supported by National Science Foundation grants DMS-0405543 and DMS-9971586.

method of smoothing results in a better estimator of the dissimilarities among a set of noisy curves. For a model having independent noise structure, the smoothed-data dissimilarity estimator dominates the observed-data estimator. For a dependent-error model—often applicable when the functional data are measured nearly continuously over some domain—an asymptotic domination result is given for the smoothed-data estimator. A simulation study indicates the magnitude of improvement provided by the shrinkage estimator and examines its behavior for heavy-tailed noise structure.

The shrinkage estimator presented here combines Stein estimation and basis function-based linear smoothers in a novel manner. Statisticians increasingly analyze sizable sets of functional data, and the results in this paper are a useful contribution to the theory of the effect of pre-smoothing on functional data analysis.
KEY WORDS: Distance methods; Dissimilarity measures; Cluster analysis; Multidimensional scaling; Smoothing; Statistical matching; Stein estimation; Shrinkage estimation.

1 Introduction

Measures of dissimilarity, or distance, among objects in a data set are fundamental to a number of statistical methods. Primary among these is cluster analysis (Everitt, Landau and Leese 2001; Kaufman and Rousseeuw 1990), but other methods such as multidimensional scaling (Young and Hamer 1987) and statistical matching (Rodgers 1988) are typically based on pairwise dissimilarities among data (Johnson and Wichern 1998, chap. 12).

If the measurements on the objects are multivariate and continuous, Euclidean distance is a popular dissimilarity metric, while other types of data require specialized dissimilarity measures. In general, the dissimilarities among N objects can be summarized with an $N \times N$ symmetric dissimilarity matrix, whose (i, j) entry is the dissimilarity between object i and object j.

If the observed data have random variation, and hence the measurements on the objects contain error, then the distances between pairs of objects will have error. Consider the problem of clustering objects in a data set using some standard algorithm. If we want our algorithm to produce a clustering result that is close to the underlying structure, it seems desirable that the dissimilarity matrix for the data we use reflect as closely as possible the (unknown) pairwise dissimilarities between the underlying systematic components of the data. A small simulation illustrates the intuitive notion that if the dissimilarities in the observed distance matrix are near the "truth," then the resulting clustering structure should be near the true structure.

We generate a sample of 60 3-dimensional normal random variables (with covariance matrix $\mathbf{I}$) such that 15 observations have mean vector $(1, 3, 1)'$, 15 have mean $(10, 6, 4)'$, 15 have mean $(1, 10, 2)'$, and 15 have mean $(5, 1, 10)'$. These means are well-separated enough that the data naturally form four clusters, and the true clustering is obvious. Then for 100 iterations we perturb the data with random $N(0, \sigma^2)$ noise having varying values of σ. For each iteration, we compute the dissimilarities and input the dissimilarity matrix of the perturbed data into the K-medoids clustering algorithm (see Kaufman and Rousseeuw 1987) and obtain a resulting clustering.

Figure 1 plots, for each perturbed data set, the mean (across elements)

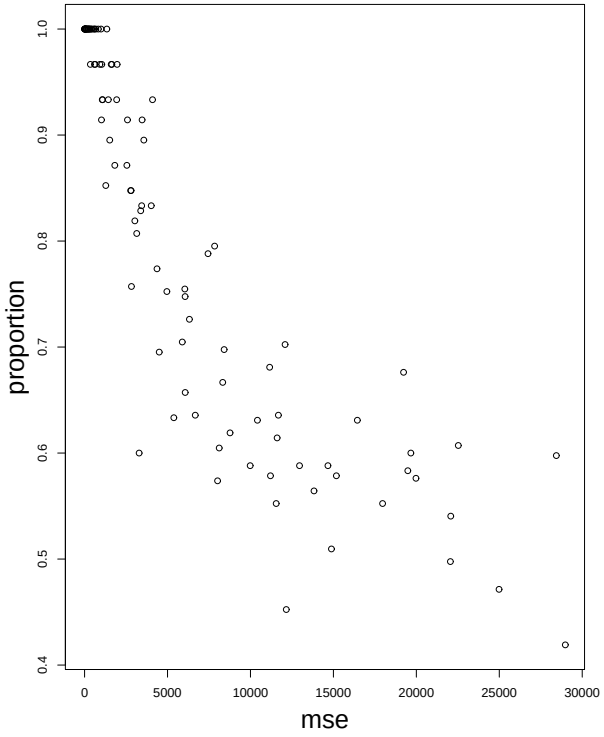

Figure 1: Proportion of pairs of objects correctly grouped vs. MSE of dissimilarities.

squared discrepancy from the true dissimilarity matrix against the proportion of all possible pairs of objects that are correctly matched in the clustering resulting from the perturbed matrix. (A correct match for two objects means correctly putting the two objects in the same cluster *or* correctly putting the two objects in different clusters, depending on the "truth.") This proportion serves as a measure of concordance between the clustering of the perturbed data set and the underlying clustering structure. We expect that as mean squared discrepancy among dissimilarities increases, the proportion of pairs correctly clustered will decrease, and the plot indicates this negative association. This indicates that a better estimate of the pairwise dissimilarities among the data tends to yield a better estimate of the true clustering structure.

We propose that, when encountering noisy data, it is advantageous to use not the observed data, but rather a pre-smoothed version of the data, to estimate the dissimilarities. We focus in particular on the case of functional data, in which the underlying process generating the data is a smooth, continuous curve. We propose a natural dissimilarity measure for data sets of (discretized) curves, and we show that a James-Stein-type smoothing method can yield a dissimilarity estimator that dominates the usual estimator based on no smoothing.

The paper is organized as follows. In Section 2, we discuss two possible models, and a dissimilarity measure, for noisy functional data. Section 3.1 contains a domination result for the James-Stein dissimilarity estimator under a "discrete noise" model, which is extended in Section 3.2. In Section 4, we give an asymptotic domination result for a "functional noise" model. In

Section 5, we present a simulation study which indicates the magnitude of improvement provided by the James-Stein estimator and examines its behavior for heavy-tailed noise structure. Section 6 is a discussion, and various proofs and technical details are stated in the appendices.

2 A Dissimilarity Measure for Functional Data

Frequently, the measurements on each observation are connected by being part of a single underlying continuous process (often, but not always, a time process). One example of such data are the growth records of Swiss boys (Falkner 1960), discussed by Ramsay and Silverman (1997, p. 2), in which the measurements are the heights of the boys at 29 different ages. Ramsay and Silverman (1997) generally label such data as *functional data,* since the underlying data are thought to be intrinsically smooth, continuous curves having domain $\mathcal{T}$, which without loss of generality we take to be $[0, T]$. The observed data vector $\mathbf{y}$ is merely a discretized representation of the functional observation $y(t)$. Typically, in functional data analysis (a term attributed to Ramsay and Dalzell (1991)), the primary goal is to discover something about the smooth curves that underlie the functional observations, and to analyze the entire set of functional data (consisting of many curves).

When scientists observe data containing random noise, they typically desire to remove the random variation to better understand the underlying process of interest. Often, when functional data are analyzed, the vector of measurements is converted to a curve via a smoothing procedure which reduces the random variation in the function. Scatterplot smoothing, or non-

parametric regression, may be used generally for paired data (t_i, y_i) for which some underlying regression function $E[y_i] = f(t_i)$ is assumed. But smoothing is particularly appropriate for functional data, for which a functional relationship $y(t)$ between the response and the process on $\mathcal{T}$ is inherent in the data.

We denote the "observed" noisy curves by $y_1(t), \ldots, y_N(t)$, and their underlying signal curves by $\mu_1(t), \ldots, \mu_N(t)$. In reality we observe these curves at a grid of n points, $t_1, \ldots, t_n$, so that we observe N independent vectors, each $n \times 1$: $\mathbf{y}_1, \ldots, \mathbf{y}_N$.

A possible model for our noisy data is the *discrete noise model*:

$$\mathbf{y}_{ij} = \mu_i(t_j) + \boldsymbol{\epsilon}_{ij}, i = 1, \ldots, N, j = 1, \ldots, n. \tag{1}$$

Here, for each $i = 1, \ldots, N$, $\boldsymbol{\epsilon}_{ij}$ may be considered independent for different measurement points, having a normal distribution with mean zero and constant variance σ_i^2.

Another possible model for our noisy curves is the *functional noise model*:

$$y_i(t_j) = \mu_i(t_j) + \epsilon_i(t_j), i = 1, \ldots, N, j = 1, \ldots, n, \tag{2}$$

where $\epsilon_i(t)$ is, for example, a stationary Ornstein-Uhlenbeck process with "pull" parameter $\beta > 0$ and variability parameter σ_i^2. This choice of model implies that the errors for the ith discretized curve have variance-covariance matrix $\boldsymbol{\Sigma}_i = \sigma_i^2 \boldsymbol{\Omega}$ where $\boldsymbol{\Omega}_{lm} = (2\beta)^{-1} \exp(-\beta |t_l - t_m|)$ (Taylor, Cumberland and Sy 1994). Note that in this case, the noise process is functional—specifically Ornstein-Uhlenbeck—but we still assume the response data collected is discretized, and is thus a vector at the level of analysis. Concep-

tually, however, the noise process is smooth and continuous in (2), as is the signal process in either model (1) or (2).

Depending on the data and sampling scheme, either (1) or (2) may be an appropriate model. If the randomness in the data arises from measurement error that is independent from one measurement to the next, (1) is more appropriate. Ramsay and Silverman (1997, p. 42) suggest a discrete noise model for the Swiss growth data, in which heights of boys are measured at 29 separate ages, and in which some small measuring error (independent across measurements) is likely to be present in the recorded data.

In the case that the variation of the observed data from the underlying curve is due to an essentially continuous random process, model (2) may be appropriate. Data that are measured frequently and *almost* continuously—for example, via sophisticated monitoring equipment—may be more likely to follow model (2), since data measured closely across time (or another domain) may more likely be correlated. We will examine both situations.

In practice, we apply a smoothing matrix $\mathbf{S}$ to the observed noisy data to obtain a smooth, called *linear* when the smooth $\hat{\boldsymbol{\mu}}_{\boldsymbol{i}}$ can be written as

$$\hat{\boldsymbol{\mu}}_{\boldsymbol{i}} = \mathbf{S}\mathbf{y}_i, i = 1, \ldots, N$$

where $\mathbf{S}$ does not depend on $\mathbf{y}_i$ (Buja, Hastie and Tibshirani 1989). Let $\hat{\mu}_i(t)$ be the smooth corresponding to the signal curve $\mu_i(t)$, for $i = 1, \ldots, n$, and then $\hat{\boldsymbol{\mu}}_{\boldsymbol{i}} = (\hat{\mu}_i(t_1), \ldots, \hat{\mu}_i(t_n))'$. Note that as $n \to \infty$, the vector $\hat{\boldsymbol{\mu}}_{\boldsymbol{i}}$ begins to closely resemble the curve $\hat{\mu}_i(t)$ on $[0, T]$.

(Here and subsequently, when writing "limit as $n \to \infty$," we assume $t_1, \ldots, t_n \in [0, T]$; that is, the collection of points is becoming denser within $[0, T]$, with the maximum gap between any pair of adjacent points $t_{i-1}, t_i, i =$

$2, \ldots, n$, tending to 0. Stein (1995) calls this method of taking the limit "fixed-domain asymptotics," while Cressie (1993) calls it "infill asymptotics.")

Many popular smoothing methods (kernel smoothers, local polynomial regression, smoothing splines) are linear. Note that if a bandwidth or smoothing parameter for these methods is chosen via a data-driven method, then technically, these smoothers become nonlinear (Buja et al. 1989).

We will focus primarily on basis function smoothing methods, in which the smoothing matrix $\mathbf{S}$ is an orthogonal projection (i.e., symmetric and idempotent). For a linear basis function smoother that is fitted via least squares, $\mathbf{S}$ will be symmetric and idempotent as long as the n points at which $\hat{\boldsymbol{\mu}}_i$ is evaluated are identical to the points at which $\mathbf{y}_i$ is observed (Ramsay and Silverman 1997, p. 44). These methods seek to express the signal curve as a linear combination of k ($< n$) specified basis functions. Assuming the matrix of these basis functions evaluated at $t_1, \ldots, t_n$ has full column rank, the rank of $\mathbf{S}$ is k. Examples of such smoothers are regression splines (in particular, B-splines), some wavelet bases, and Fourier series bases (Ramsay and Silverman 1997). Regression splines and B-spline bases are discussed in detail by de Boor (1978) and Eubank (1988, chap. 7).

If we choose squared L_2 distance as our dissimilarity metric, then denote the dissimilarities between the true, observed, and smoothed curves i and j, respectively, as follows:

$$\delta_{ij} = \int_0^T [\mu_i(t) - \mu_j(t)]^2 \, dt, \tag{3}$$

$$\hat{\delta}_{ij} = \int_0^T [y_i(t) - y_j(t)]^2 \, dt, \tag{4}$$

$$\hat{\delta}_{ij}^{(smooth)} = \int_0^T [\hat{\mu}_i(t) - \hat{\mu}_j(t)]^2 \, dt. \tag{5}$$

Define $\boldsymbol{\theta_{ij}} = \boldsymbol{\mu_i} - \boldsymbol{\mu_j}$ where $\boldsymbol{\mu_i} = (\mu_i(t_1), \ldots, \mu_i(t_n))'$; $\hat{\boldsymbol{\theta}}_{\boldsymbol{ij}} = \mathbf{y}_i - \mathbf{y}_j$; and note $\mathbf{S}\hat{\boldsymbol{\theta}}_{\boldsymbol{ij}} = \hat{\boldsymbol{\mu}}_{\boldsymbol{i}} - \hat{\boldsymbol{\mu}}_{\boldsymbol{j}}$. Then, if the data follow the discrete noise model, $\hat{\boldsymbol{\theta}}_{\boldsymbol{ij}} \sim N(\boldsymbol{\theta_{ij}}, \sigma_{ij}^2 \mathbf{I})$ where $\sigma_{ij}^2 = (\sigma_i^2 + \sigma_j^2)$. If the data follow the functional noise model, $\hat{\boldsymbol{\theta}}_{\boldsymbol{ij}} \sim N(\boldsymbol{\theta_{ij}}, \boldsymbol{\Sigma}_{ij})$ where $\boldsymbol{\Sigma}_{ij} = \sigma_{ij}^2 \boldsymbol{\Omega}$ and $\sigma_{ij}^2 = (\sigma_i^2 + \sigma_j^2)$.

If we observe the response at points $t_1, \ldots, t_n$ in $[0, T]$, then we may approximate (3)-(5) by

$$d_{ij} = \frac{T}{n} \boldsymbol{\theta}_{\boldsymbol{ij}}' \boldsymbol{\theta_{ij}}, \quad \hat{d}_{ij} = \frac{T}{n} \hat{\boldsymbol{\theta}}_{\boldsymbol{ij}}' \hat{\boldsymbol{\theta}}_{\boldsymbol{ij}}, \quad \hat{d}_{ij}^{(smooth)} = \frac{T}{n} \hat{\boldsymbol{\theta}}_{\boldsymbol{ij}}' \mathbf{S}' \mathbf{S} \hat{\boldsymbol{\theta}}_{\boldsymbol{ij}}.$$

The question of interest is: When, for large n, is $\hat{d}_{ij}^{(smooth)}$ a better estimator of d_{ij} than is $\hat{d}_{ij}$?

(Note: In the following sections, since the pair of curves i and j is arbitrary, we shall suppress the ij subscript on $\hat{\boldsymbol{\theta}}_{\boldsymbol{ij}}$, $\boldsymbol{\theta_{ij}}$, σ_{ij}^2, and $\boldsymbol{\Sigma}_{ij}$, writing instead $\hat{\boldsymbol{\theta}}$, $\boldsymbol{\theta}$, σ^2, and $\boldsymbol{\Sigma}$, understanding that we are concerned with any particular pair $i, j \in \{1, \ldots, N\}, i \neq j$.)

3 Case I: Data Following the Discrete Noise Model

First we will consider functional data following model (1). Recall that we assume the response is measured at n discrete points in $[0, T]$.

3.1 Dissimilarity Estimation for Known σ^2

We assume that $\mathbf{S}$ is symmetric and idempotent and projects the observed data onto a lower-dimensional space (of dimension $k < n$), and thus $r(\mathbf{S}) = tr(\mathbf{S}) = k$, where $r(\cdot)$ denotes rank and $tr(\cdot)$ denotes trace. Note that $\mathbf{S}$ is a *shrinking smoother,* since all its singular values are in $[0, 1]$ (Buja et al. 1989). Recall that according to the discrete noise model for the data, $\hat{\boldsymbol{\theta}} \sim N(\boldsymbol{\theta}, \sigma^2\mathbf{I})$, with σ^2 assumed known in this section. Without loss of generality, let $\sigma^2\mathbf{I} = \mathbf{I}$. (Otherwise, we can let, for example, $\hat{\boldsymbol{\eta}} = \sigma^{-1}\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\eta} = \sigma^{-1}\boldsymbol{\theta}$ and work with $\hat{\boldsymbol{\eta}}$ and $\boldsymbol{\eta}$ instead.)

Recall that $\frac{T}{n}\hat{\boldsymbol{\theta}}'\hat{\boldsymbol{\theta}}$ represents the approximate L_2 distance between observed curves $y_i(t)$ and $y_j(t)$ and $\frac{T}{n}\hat{\boldsymbol{\theta}}'\mathbf{S}'\mathbf{S}\hat{\boldsymbol{\theta}} = \frac{T}{n}\hat{\boldsymbol{\theta}}'\mathbf{S}\hat{\boldsymbol{\theta}}$ represents the approximate L_2 distance between smoothed curves $\hat{\mu}_i(t)$ and $\hat{\mu}_j(t)$.

We may determine when the "smoothed-data dissimilarity" better estimates the true dissimilarity δ_{ij} between curves $\mu_i(t)$ and $\mu_j(t)$ than the observed-data dissimilarity by comparing the risks of the two estimators. (Recall that the *risk* of an estimator $\hat{\tau}$ for τ, given by $R(\tau, \hat{\tau}) = E[L(\tau, \hat{\tau})]$ is, for the familiar case of squared error loss $L(\tau, \hat{\tau}) = (\tau - \hat{\tau})^2$, simply the mean squared error (MSE) of the estimator.) Hence, we propose to compare the MSEs of two competing estimators and choose the one with the smaller MSE.

First, consider the case in which $\boldsymbol{\theta}$ lies in the linear subspace that $\mathbf{S}$ projects onto, i.e., $\mathbf{S}\boldsymbol{\theta} = \boldsymbol{\theta}$. Note that if two arbitrary (discretized) signal curves $\boldsymbol{\mu_i}$ and $\boldsymbol{\mu_j}$ are in this linear subspace, then the corresponding $\boldsymbol{\theta}$ is also in the subspace, since in this case

$$\boldsymbol{\theta} = \boldsymbol{\mu_i} - \boldsymbol{\mu_j} = \mathbf{S}\boldsymbol{\mu_i} - \mathbf{S}\boldsymbol{\mu_j} = \mathbf{S}(\boldsymbol{\mu_i} - \boldsymbol{\mu_j}) = \mathbf{S}\boldsymbol{\theta}.$$

In this idealized situation, a straightforward comparison of MSEs shows that the smoothed-data estimator improves on the observed-data estimator.

If the smooth $\mathbf{S}\boldsymbol{\theta} \neq \boldsymbol{\theta}$, it can be shown that some shrinkage smoothing of the observed curves makes the dissimilarity estimator better, but too much shrinkage leads to a forfeiture of that advantage. The disadvantage of the linear smoother is that it cannot "learn" from the data how much to shrink $\hat{\boldsymbol{\theta}}$. To improve the smoother, we can employ a James-Stein-type adjustment to $\mathbf{S}$, so that the data can determine the amount of shrinkage.

What is now known as "shrinkage estimation" or "Stein estimation" originated with the work of Stein in the context of estimating a multivariate normal mean. In subsequent years, many results have been derived about shrinkage estimation in a variety of contexts. As part of a detailed discussion of shrinkage estimation, Lehmann and Casella (1998, p. 367) discuss shrinking an estimator toward a linear subspace of the parameter space. For example, Casella and Hwang (1987) propose such shrinkage estimators in the context of confidence sets for a multivariate normal mean. Green and Strawderman (1991), also in the context of estimating a multivariate mean, discuss how shrinking an unbiased estimator toward a possibly biased estimator using a James-Stein form can result in a risk improvement.

In our case, we believe that $\boldsymbol{\theta}$ is near $\mathbf{S}\boldsymbol{\theta}$, so we shrink $\hat{\boldsymbol{\theta}}$ toward $\mathbf{S}\hat{\boldsymbol{\theta}}$, obtaining a James-Stein estimator of $\boldsymbol{\theta}$ (see Lehmann and Casella, 1998, p. 367):

$$\hat{\boldsymbol{\theta}}^{(JS)} = \mathbf{S}\hat{\boldsymbol{\theta}} + \left(1 - \frac{a}{||\hat{\boldsymbol{\theta}} - \mathbf{S}\hat{\boldsymbol{\theta}}||^2}\right)(\hat{\boldsymbol{\theta}} - \mathbf{S}\hat{\boldsymbol{\theta}}) \tag{6}$$

where a is a constant and $||\cdot||$ is the usual Euclidean norm. In practice, to avoid the problem of the shrinkage factor possibly being negative for small

$||\hat{\boldsymbol{\theta}} - \mathbf{S}\hat{\boldsymbol{\theta}}||^2$, we would use the positive-part James-Stein estimator

$$\hat{\boldsymbol{\theta}}_+^{(\boldsymbol{JS})} = \mathbf{S}\hat{\boldsymbol{\theta}} + \left(1 - \frac{a}{||\hat{\boldsymbol{\theta}} - \mathbf{S}\hat{\boldsymbol{\theta}}||^2}\right)_+ (\hat{\boldsymbol{\theta}} - \mathbf{S}\hat{\boldsymbol{\theta}}) \tag{7}$$

where $x_+ = xI(x \geq 0)$.

The shrinkage estimator involves the data by giving more weight to $\hat{\boldsymbol{\theta}}$ when $||\hat{\boldsymbol{\theta}} - \mathbf{S}\hat{\boldsymbol{\theta}}||^2$ is large and more weight to $\mathbf{S}\hat{\boldsymbol{\theta}}$ when $||\hat{\boldsymbol{\theta}} - \mathbf{S}\hat{\boldsymbol{\theta}}||^2$ is small. In fact, if the smoother is at all well-chosen, $\mathbf{S}\hat{\boldsymbol{\theta}}$ is often close enough to $\hat{\boldsymbol{\theta}}$ that the shrinkage factor in (7) is very often zero. The shrinkage factor is actually merely a safeguard against oversmoothing, in case $\mathbf{S}$ smooths the curves beyond what, in reality, it should.

Let us consider an appropriate shrinkage estimator of $d_{ij} = \frac{T}{n}\boldsymbol{\theta}'\boldsymbol{\theta}$, namely:

$$\hat{d}_{ij}^{(JS)} = \frac{T}{n}\hat{\boldsymbol{\theta}}^{(\boldsymbol{JS})\prime}\hat{\boldsymbol{\theta}}^{(\boldsymbol{JS})} \tag{8}$$

where $\hat{\boldsymbol{\theta}}^{(\boldsymbol{JS})}$ is given by (6).

Theorem 1 *Suppose the observed* $\hat{\boldsymbol{\theta}} \sim N(\boldsymbol{\theta}, \sigma^2\mathbf{I})$. *Let* $\mathbf{S}$ *be a symmetric and idempotent linear smoothing matrix of rank* k. *If* $n - k > 4$, *then there exists a positive real number* r *such that for* $0 < a < r$, *the risk difference is negative and the smoothed-data dissimilarity estimator* $\hat{d}_{ij}^{(JS)}$ *has smaller risk than* $\hat{d}_{ij}$.

Proof of Theorem 1: The risk difference between the James-Stein smoothed dissimilarity $\hat{d}_{ij}^{(JS)}$ and the observed dissimilarity $\hat{d}_{ij}$ is:

$$\frac{T^2}{n^2}\Delta = E\left[\left(\frac{T}{n}\hat{\boldsymbol{\theta}}^{(\boldsymbol{JS})\prime}\hat{\boldsymbol{\theta}}^{(\boldsymbol{JS})} - \frac{T}{n}\boldsymbol{\theta}'\boldsymbol{\theta}\right)^2\right] - E\left[\left(\frac{T}{n}\hat{\boldsymbol{\theta}}'\hat{\boldsymbol{\theta}} - \frac{T}{n}\boldsymbol{\theta}'\boldsymbol{\theta}\right)^2\right]. \tag{9}$$

Since we are interested in when the risk difference is negative, we can ignore the positive constant multiplier T^2/n^2.

From (9),

$$\Delta = E[(\hat{\boldsymbol{\theta}}^{(JS)\prime}\hat{\boldsymbol{\theta}}^{(JS)})^2 - (\hat{\boldsymbol{\theta}}'\hat{\boldsymbol{\theta}})^2 - 2\boldsymbol{\theta}'\boldsymbol{\theta}(\hat{\boldsymbol{\theta}}^{(JS)\prime}\hat{\boldsymbol{\theta}}^{(JS)} - \hat{\boldsymbol{\theta}}'\hat{\boldsymbol{\theta}})].$$

Let $\hat{\boldsymbol{\theta}}^{(JS)} = \hat{\boldsymbol{\theta}} - \phi(\hat{\boldsymbol{\theta}})\hat{\boldsymbol{\theta}}$, where

$$\phi(\hat{\boldsymbol{\theta}}) = \frac{a}{||\hat{\boldsymbol{\theta}} - \mathbf{S}\hat{\boldsymbol{\theta}}||^2}(\mathbf{I} - \mathbf{S}) = \frac{a}{\hat{\boldsymbol{\theta}}'(\mathbf{I} - \mathbf{S})\hat{\boldsymbol{\theta}}}(\mathbf{I} - \mathbf{S}).$$

Then routine calculations show that $\hat{\boldsymbol{\theta}}^{(JS)\prime}\hat{\boldsymbol{\theta}}^{(JS)} - \hat{\boldsymbol{\theta}}'\hat{\boldsymbol{\theta}}$ can be written as:

$$-2\hat{\boldsymbol{\theta}}'\phi(\hat{\boldsymbol{\theta}})'\hat{\boldsymbol{\theta}} + \hat{\boldsymbol{\theta}}'\phi(\hat{\boldsymbol{\theta}})'\phi(\hat{\boldsymbol{\theta}})\hat{\boldsymbol{\theta}} = -2a + \frac{a^2}{||\hat{\boldsymbol{\theta}} - \mathbf{S}\hat{\boldsymbol{\theta}}||^2}.$$

Hence the (scaled) risk difference is

$$\Delta = E\left[\left(\hat{\boldsymbol{\theta}}'\hat{\boldsymbol{\theta}} - 2a + \frac{a^2}{||\hat{\boldsymbol{\theta}} - \mathbf{S}\hat{\boldsymbol{\theta}}||^2}\right)^2 - (\hat{\boldsymbol{\theta}}'\hat{\boldsymbol{\theta}})^2 - 2\boldsymbol{\theta}'\boldsymbol{\theta}\left(-2a + \frac{a^2}{||\hat{\boldsymbol{\theta}} - \mathbf{S}\hat{\boldsymbol{\theta}}||^2}\right)\right].$$

Note that $\hat{\boldsymbol{\theta}}'\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}'(\mathbf{I} - \mathbf{S})\hat{\boldsymbol{\theta}} + \hat{\boldsymbol{\theta}}'\mathbf{S}\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}'\boldsymbol{\theta} = \boldsymbol{\theta}'(\mathbf{I} - \mathbf{S})\boldsymbol{\theta} + \boldsymbol{\theta}'\mathbf{S}\boldsymbol{\theta}$. Define the following quadratic forms: $\hat{q}_1 = \hat{\boldsymbol{\theta}}'(\mathbf{I} - \mathbf{S})\hat{\boldsymbol{\theta}}$, $\hat{q}_2 = \hat{\boldsymbol{\theta}}'\mathbf{S}\hat{\boldsymbol{\theta}}$, $q_1 = \boldsymbol{\theta}'(\mathbf{I} - \mathbf{S})\boldsymbol{\theta}$, and $q_2 = \boldsymbol{\theta}'\mathbf{S}\boldsymbol{\theta}$.

Note that $\hat{q}_1 \sim \chi^2_{n-k}(q_1)$, $\hat{q}_2 \sim \chi^2_k(q_2)$ and they are independent (since $\hat{\boldsymbol{\theta}} \sim N(\boldsymbol{\theta}, \mathbf{I})$, $\mathbf{S}$ is idempotent, and $(\mathbf{I} - \mathbf{S})\mathbf{S} = \mathbf{0}$). Now write Δ as:

$$\begin{aligned}
\Delta &= E\left[\left([\hat{q}_1 + \hat{q}_2] - 2a + \frac{a^2}{\hat{q}_1}\right)^2 - [\hat{q}_1 + \hat{q}_2]^2 - 2(q_1 + q_2)\left(-2a + \frac{a^2}{\hat{q}_1}\right)\right] \\
&= E\left[2\left(-2a + \frac{a^2}{\hat{q}_1}\right)[\hat{q}_1 + \hat{q}_2] + \left(-2a + \frac{a^2}{\hat{q}_1}\right)^2 + 4a(q_1 + q_2) - \frac{2a^2}{\hat{q}_1}(q_1 + q_2)\right] \\
&= E\left[-4a[\hat{q}_1 + \hat{q}_2] + 4a^2 + \left(\frac{a^2}{\hat{q}_1}\right)^2 - \frac{4a^3}{\hat{q}_1} + 4a(q_1 + q_2)\right. \\
&\quad \left. + \frac{2a^2}{\hat{q}_1}(\hat{q}_1 + \hat{q}_2 - q_1 - q_2)\right].
\end{aligned}$$

Since $E[\hat{q}_1 + \hat{q}_2] = n + q_1 + q_2$,

$$\Delta = -4an + 4a^2 + E\left[\left(\frac{a^2}{\hat{q}_1}\right)^2 - \frac{4a^3}{\hat{q}_1} + \frac{2a^2}{\hat{q}_1}(\hat{q}_1 + \hat{q}_2 - q_1 - q_2)\right].$$

Since $\hat{q}_1$ and $\hat{q}_2$ are independent, we can take the expectation of $\hat{q}_2$ in the last term. Since $E[\hat{q}_2] = k + q_2$,

$$\Delta = 4a^2 - 4an + E\left[\frac{a^4}{\hat{q}_1^2} + \frac{2a^2k}{\hat{q}_1} + 2a^2\left(1 - \frac{2a + q_1}{\hat{q}_1}\right)\right].$$

By Jensen's Inequality, $E(1/\hat{q}_1) \geq 1/E(\hat{q}_1)$, and since $E(\hat{q}_1) = n - k + q_1$, we can bound the last term above:

$$E\left[2a^2\left(1 - \frac{2a + q_1}{\hat{q}_1}\right)\right] \leq 2a^2\left(1 - \frac{2a + q_1}{n - k + q_1}\right) = 2a^2\left(\frac{n - k + q_1 - 2a - q_1}{n - k + q_1}\right)$$

$$= 2a^2\left(\frac{n - k - 2a}{n - k + q_1}\right) \leq 2a^2\left(\frac{n - k - 2a}{n - k}\right) = 2a^2\left(1 - \frac{2a}{n - k}\right).$$

Hence

$$\Delta \leq 4a^2 - 4an + E\left[\frac{a^4}{\hat{q}_1^2}\right] + E\left[\frac{2a^2k}{\hat{q}_1}\right] + 2a^2\left(1 - \frac{2a}{n - k}\right). \qquad (10)$$

The numerators of the terms in the expected values in (10) are positive. The random variable $\hat{q}_1$ is a noncentral χ^2_{n-k} with noncentrality parameter q_1, and this distribution is stochastically increasing in q_1. This implies that for $m > 0$, $E[(\hat{q}_1)^{-m}]$ is decreasing in q_1. So

$$E[(\hat{q}_1)^{-m}] \leq E[(\chi^2_{n-k})^{-m}]$$

where χ^2_{n-k} is a central χ^2 with $n - k$ degrees of freedom. So by replacing $\hat{q}_1$ with χ^2_{n-k} in (10), we obtain an upper bound Δ_U for Δ:

$$\Delta_U = 4a^2 - 4an + E\left[\frac{a^4}{(\chi^2_{n-k})^2}\right] + E\left[\frac{2a^2k}{\chi^2_{n-k}}\right] + 2a^2\left(1 - \frac{2a}{n - k}\right). \qquad (11)$$

Note that $E[(\chi^2_{n-k})^{-1}] = 1/(n-k-2)$ and $E[(\chi^2_{n-k})^{-2}] = 1/(n-k-2)(n-k-4)$. So taking expectations and writing the upper bound as a function of a:

$$\Delta_U(a) = \frac{1}{(n-k-2)(n-k-4)}a^4 - \frac{4}{n-k}a^3 + \left(\frac{2k}{n-k-2} + 6\right)a^2 - 4na$$

We seek values of a that make Δ_U negative. The fourth-degree equation $\Delta_U(a) = 0$ can be solved analytically. (Two of the roots are imaginary.) One real root is clearly 0; call the second real root r. Write $\Delta_U(a) = 0$ as $c_4a^4 + c_3a^3 + c_2a^2 + c_1a = 0$, where c_4, c_3, c_2, c_1 are the respective coefficients of $\Delta_U(a)$. It is clear that if $n - k > 4$, then $c_4 > 0, c_3 < 0, c_2 > 0, c_1 < 0$.

Note that r also solves the cubic equation $f_c(a) = c_4a^3 + c_3a^2 + c_2a + c_1 = 0$. Since its leading coefficient $c_4 > 0$,

$$\lim_{a\to-\infty} f_c(a) = -\infty, \lim_{a\to\infty} f_c(a) = \infty.$$

Since $f_c(a)$ has only one real root, it crosses the a-axis only once. Since its vertical intercept $c_1 < 0$, its horizontal intercept $r > 0$.

And since $\Delta_U(a)$ has leading coefficient $c_4 > 0$,

$$\lim_{a\to\pm\infty} \Delta_U(a) = \infty.$$

Since it tends to ∞ at its endpoints, $\Delta_U(a)$ must be negative between its two real roots 0 and r. Therefore $\Delta < 0$ for $0 < a < r$. □

Using a symbolic algebra software program (such as Maple or Mathematica), one can easily obtain the formula for the second real root for general n and k and verify that the other two roots are imaginary.

Figure 2 shows Δ_U plotted as a function of a for varying n and $k = 5$. For various choices of n (and $k = 5$) Table 1 provides values of r, as well as

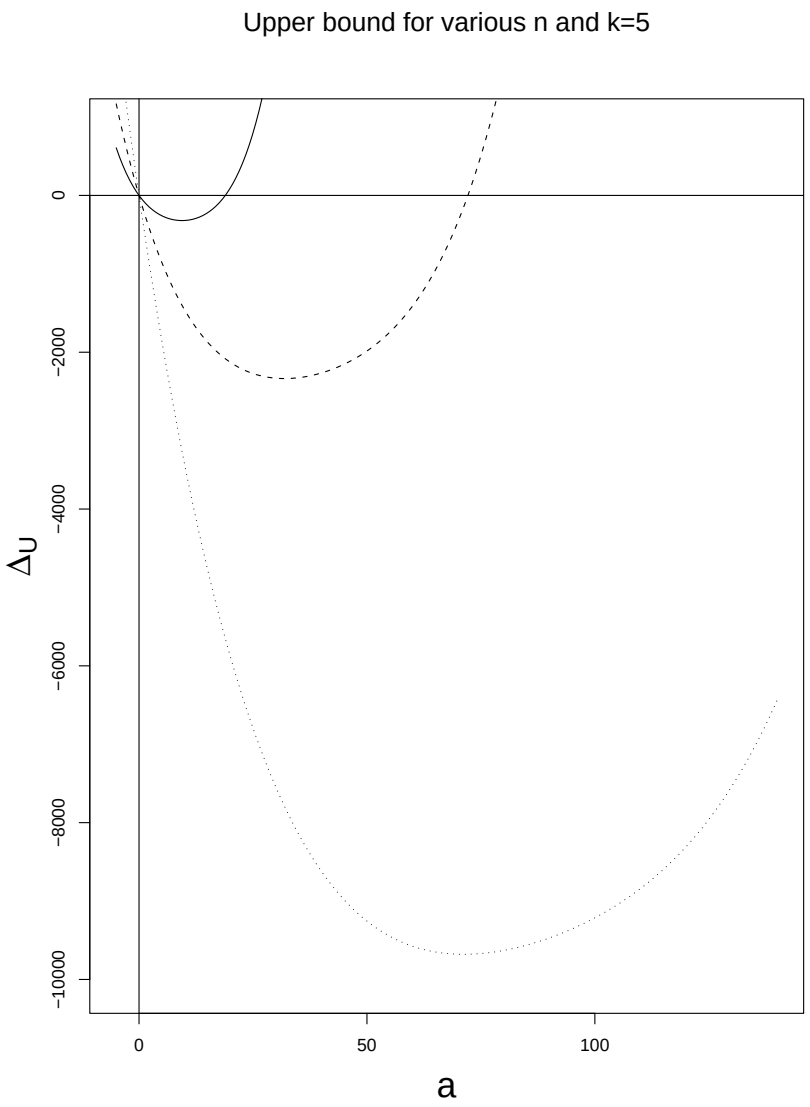

Figure 2: Plot of Δ_U against a for varying n and for $k = 5$.
Solid line: $n = 20$. Dashed line: $n = 50$. Dotted line: $n = 100$.

Table 1: Choices of a for various n and k.

n	k	**minimizer** a^*	**root** r
20	5	9.3	19.0
50	5	31.9	72.2
100	5	71.3	169.1
200	5	153.2	367.5

the value of a that minimizes the upper bound for Δ. For $0 < a < r$, the risk difference is assured of being negative. For a^*, Δ_U is minimized.

Since Δ_U provides an upper bound for the (scaled) risk difference, it may be valuable to ascertain the size of the discrepancy between Δ_U and Δ. We can estimate this discrepancy via Monte Carlo simulation. We generate a large number of random variables having the distribution of $\hat{q}_1$ (namely $\chi^2_{n-k}(q_1)$) and get an estimate of Δ using a Monte Carlo mean. For various values of q_1, Figure 3 shows Δ plotted alongside Δ_U.

3.2 Extension to Unknown σ^2

In the previous section, σ^2 was assumed to be known. We now examine the situation in which $\hat{\boldsymbol{\theta}} \sim N(\boldsymbol{\theta}, \sigma^2\mathbf{I})$ with σ^2 unknown. We suppose there exists a random variable S^2, independent of $\hat{\boldsymbol{\theta}}$, such that $S^2/\sigma^2 \sim \chi^2_\nu$. Then let $\hat{\sigma}^2 = S^2/\nu$.

Consider the following definition of the James-Stein estimator which accounts for σ^2:

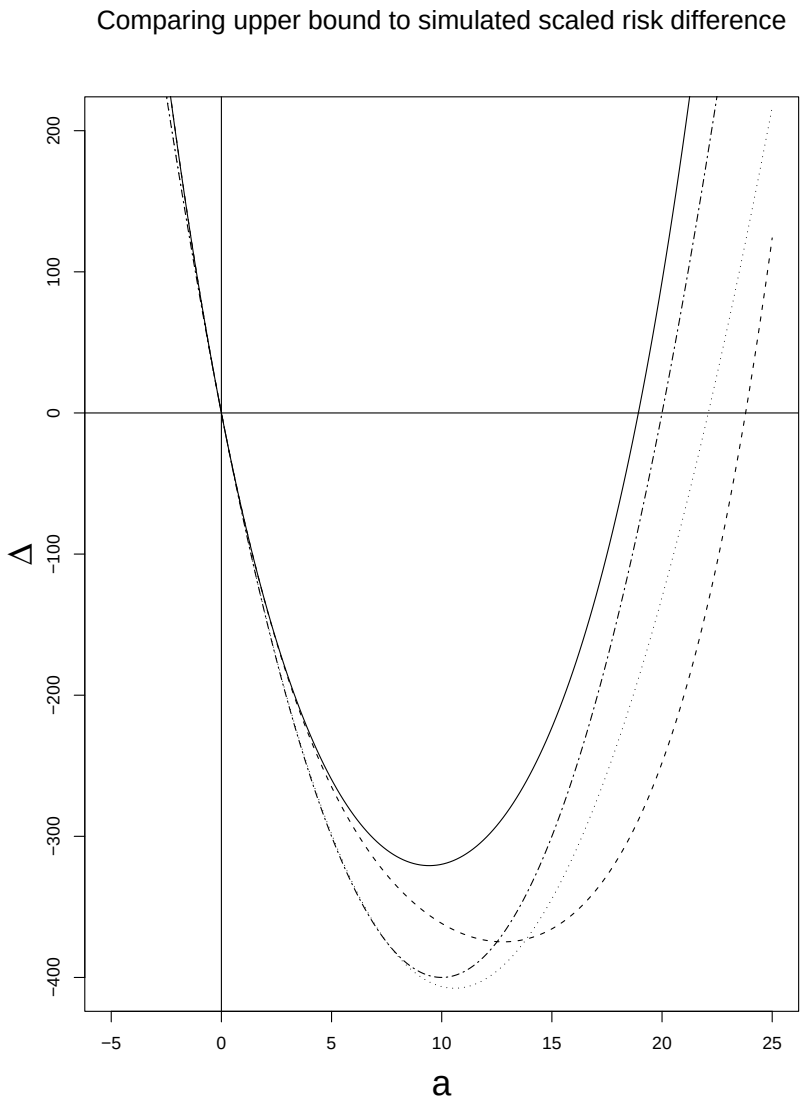

Figure 3: Plot of simulated Δ and Δ_U against a for $n = 20$, $k = 5$. Solid line: Upper bound Δ_U. Dashed line: simulated Δ, $q_1 = 0$. Dotted line: simulated Δ, $q_1 = 135$. Dot-dashed line: simulated Δ, $q_1 = 3750000$.

$$\hat{\boldsymbol{\theta}}^{(\boldsymbol{JS})} = \mathbf{S}\hat{\boldsymbol{\theta}} + \left(1 - \frac{a\sigma^2}{||\hat{\boldsymbol{\theta}} - \mathbf{S}\hat{\boldsymbol{\theta}}||^2}\right)(\mathbf{I} - \mathbf{S})\hat{\boldsymbol{\theta}}. \tag{12}$$

Note that the James-Stein estimator from Section 3 (when we assumed a known covariance matrix $\mathbf{I}$) is simply (12) with $\sigma^2 = 1$.

Now, replacing σ^2 with the estimate $\hat{\sigma}^2$, define

$$\hat{\boldsymbol{\theta}}^{(\boldsymbol{JS})}_{\hat{\boldsymbol{\sigma}}} = \mathbf{S}\hat{\boldsymbol{\theta}} + \left(1 - \frac{a\hat{\sigma}^2}{\hat{q}_1}\right)(\mathbf{I} - \mathbf{S})\hat{\boldsymbol{\theta}},$$

where $\hat{q}_1 = \hat{\boldsymbol{\theta}}'(\mathbf{I} - \mathbf{S})\hat{\boldsymbol{\theta}}$ as in Section 3.1. Define the James-Stein smoothed-data dissimilarity estimator based on $\hat{\boldsymbol{\theta}}^{(\boldsymbol{JS})}_{\hat{\boldsymbol{\sigma}}}$ to be:

$$\hat{d}^{(JS)}_{ij,\hat{\sigma}} = \frac{T}{n}\hat{\boldsymbol{\theta}}^{(\boldsymbol{JS})'}_{\hat{\boldsymbol{\sigma}}}\hat{\boldsymbol{\theta}}^{(\boldsymbol{JS})}_{\hat{\boldsymbol{\sigma}}}.$$

Theorem 2 *Suppose that* $\hat{\boldsymbol{\theta}} \sim N(\boldsymbol{\theta}, \sigma^2\mathbf{I})$ *with* σ^2 *unknown and that there exists a random variable* S^2*, independent of* $\hat{\boldsymbol{\theta}}$*, such that* $S^2/\sigma^2 \sim \chi^2_\nu$*, and let* $\hat{\sigma}^2 = S^2/\nu$*. If* $n - k > 4$*, then there exists a positive real* r *such that for* $0 < a < r$*, the risk difference is negative and the smoothed-data dissimilarity estimator* $\hat{d}^{(JS)}_{ij,\hat{\sigma}}$ *has smaller risk than* $\hat{d}_{ij}$*.*

Proof of Theorem 2: Calculations similar to those at the beginning of the proof of Theorem 1 allow us to write the scaled risk difference as:

$$\Delta_{\hat{\sigma}} = E\left[-2\left(2a\hat{\sigma}^2 - \frac{a^2\hat{\sigma}^4}{\hat{q}_1}\right)(\hat{\boldsymbol{\theta}}'\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}'\boldsymbol{\theta}) + \left(2a\hat{\sigma}^2 - \frac{a^2\hat{\sigma}^4}{\hat{q}_1}\right)^2\right]. \tag{13}$$

Since $\hat{\sigma}$ and $\hat{\boldsymbol{\theta}}$ are independent,

$$E[-4a\hat{\sigma}^2(\hat{\boldsymbol{\theta}}'\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}'\boldsymbol{\theta})] = E[-4a\hat{\sigma}^2]E[\hat{\boldsymbol{\theta}}'\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}'\boldsymbol{\theta}] = E[-4a\hat{\sigma}^2]\sigma^2 n$$

and

$$\begin{aligned} & E\left[\frac{2a^2\hat{\sigma}^4}{\hat{q}_1}(\hat{\boldsymbol{\theta}}'\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}'\boldsymbol{\theta})\right] \\ = \; & E\left[\frac{2a^2\hat{\sigma}^4}{\hat{q}_1}(\hat{q}_1 + \hat{q}_2 - q_1 - q_2)\right] \\ = \; & E[2a^2\hat{\sigma}^4]E\left[\frac{\hat{q}_2}{\hat{q}_1}\right] + E\left[\frac{2a^2\hat{\sigma}^4}{\hat{q}_1}(\hat{q}_1 - q_1 - q_2)\right] \qquad (14) \end{aligned}$$

by the independence of $\hat{\sigma}$ and $\hat{\boldsymbol{\theta}}$ (and thus $\hat{\sigma}$ and $\hat{q}_2$).

Since $\hat{q}_1$ and $\hat{q}_2$ are independent, and $E[\hat{q}_2] = q_2 + \sigma^2 k$, we may write (14) as:

$$E[2a^2\hat{\sigma}^4](q_2 + \sigma^2 k)E[1/\hat{q}_1] + E\left[\frac{2a^2\hat{\sigma}^4}{\hat{q}_1}(\hat{q}_1 - q_1 - q_2)\right]. \qquad (15)$$

Now, since $a^2\hat{\sigma}^4/\hat{q}_1$ is decreasing in $\hat{q}_1$ and $\hat{q}_1 - q_1 - q_2$ is increasing in $\hat{q}_1$, these quantities have covariance ≤ 0. Hence

$$\begin{aligned} (15) \; \leq \; & E[2a^2\hat{\sigma}^4](q_2 + \sigma^2 k)E[1/\hat{q}_1] + E\left[\frac{2a^2\hat{\sigma}^4}{\hat{q}_1}\right]E[\hat{q}_1 - q_1 - q_2] \\ = \; & E[2a^2\hat{\sigma}^4](q_2 + \sigma^2 k)E[1/\hat{q}_1] + E[2a^2\hat{\sigma}^4]E[1/\hat{q}_1](\sigma^2(n-k) - q_2) \\ = \; & E[2a^2\hat{\sigma}^4]E[1/\hat{q}_1]\sigma^2 n. \end{aligned}$$

So, from (13),

$$\Delta_{\hat{\sigma}} \leq E[-4a\hat{\sigma}^2]\sigma^2 n + E[2a^2\hat{\sigma}^4]E\left[\frac{1}{\hat{q}_1}\right]\sigma^2 n + E\left[4a^2\hat{\sigma}^4 - \frac{4a^3\hat{\sigma}^6}{\hat{q}_1} + \frac{a^4\hat{\sigma}^8}{\hat{q}_1^2}\right].$$

Note that $E[a\hat{\sigma}^2] = a\sigma^2$, $E[a^2\hat{\sigma}^4] = a^2\sigma^4\nu(\nu+2)/\nu^2$, $E[a^3\hat{\sigma}^6] = a^3\sigma^6\nu(\nu+2)(\nu+4)/\nu^3$, $E[a^4\hat{\sigma}^8] = a^4\sigma^8\nu(\nu+2)(\nu+4)(\nu+6)/\nu^4$. Since $\hat{\sigma}^2$ and $\hat{q}_1$ are independent, taking expectations:

$$\begin{aligned} \Delta_{\hat{\sigma}} \; \leq \; & -4a\sigma^4 n + 2a^2\sigma^6 n\left(\frac{\nu(\nu+2)}{\nu^2}\right)E\left[\frac{1}{\hat{q}_1}\right] + 4a^2\sigma^4\left(\frac{\nu(\nu+2)}{\nu^2}\right) \\ & - 4a^3\sigma^6\left(\frac{\nu(\nu+2)(\nu+4)}{\nu^3}\right)E\left[\frac{1}{\hat{q}_1}\right] + a^4\sigma^8\left(\frac{\nu(\nu+2)(\nu+4)(\nu+6)}{\nu^4}\right)E\left[\frac{1}{\hat{q}_1^2}\right] \end{aligned}$$

Define $\hat{q}_1^* = \hat{q}_1/\sigma^2 = (\hat{\boldsymbol{\theta}}/\sigma)'(\mathbf{I}-\mathbf{S})(\hat{\boldsymbol{\theta}}/\sigma)$. Since $\hat{\boldsymbol{\theta}} \sim N(\boldsymbol{\theta}, \sigma^2\mathbf{I})$, $(\hat{\boldsymbol{\theta}}/\sigma) \sim N(\boldsymbol{\zeta}, \mathbf{I})$ where $\boldsymbol{\zeta} = \boldsymbol{\theta}/\sigma$. Then the risk difference is a function of $\boldsymbol{\zeta}$, and the distribution of $\hat{q}_1^*$ is a function of $\boldsymbol{\zeta}$ and does not involve σ alone. Now we may divide the inequality through by $\sigma^4 > 0$ to obtain

$$\begin{aligned}\frac{\Delta_{\hat{\sigma}}}{\sigma^4} \quad \leq \quad & -4an + 2a^2 n\left(\frac{\nu(\nu+2)}{\nu^2}\right)E\left[\frac{1}{\hat{q}_1^*}\right] + 4a^2\left(\frac{\nu(\nu+2)}{\nu^2}\right) \\ & - 4a^3\left(\frac{\nu(\nu+2)(\nu+4)}{\nu^3}\right)E\left[\frac{1}{\hat{q}_1^*}\right] + a^4\left(\frac{\nu(\nu+2)(\nu+4)(\nu+6)}{\nu^4}\right)E\left[\frac{1}{\hat{q}_1^{*2}}\right].\end{aligned}$$

Now, since $(\mathbf{I}-\mathbf{S})\mathbf{I}$ is idempotent, $\hat{q}_1^*$ is noncentral $\chi^2_{n-k}(\boldsymbol{\zeta}'(\mathbf{I}-\mathbf{S})\boldsymbol{\zeta})$. Recall that the noncentral χ^2 distribution is stochastically increasing in its noncentrality parameter, so we may replace this parameter by zero and obtain the upper bound

$$E[(\hat{q}_1^*)^{-m}] \leq E[(\chi^2_{n-k})^{-m}], m = 1, 2.$$

Hence

$$\begin{aligned}\frac{\Delta_{\hat{\sigma}}}{\sigma^4} \leq \Delta_U \quad = \quad & -4an + 2a^2 n\left(\frac{\nu(\nu+2)}{\nu^2}\right)E[(\chi^2_{n-k})^{-1}] + 4a^2\left(\frac{\nu(\nu+2)}{\nu^2}\right) \\ & - 4a^3\left(\frac{\nu(\nu+2)(\nu+4)}{\nu^3}\right)E[(\chi^2_{n-k})^{-1}] \\ & + a^4\left(\frac{\nu(\nu+2)(\nu+4)(\nu+6)}{\nu^4}\right)E[(\chi^2_{n-k})^{-2}]\end{aligned}$$

If $n-k > 4$, then $E[(\chi^2_{n-k})^{-1}] = 1/(n-k-2)$ and $E[(\chi^2_{n-k})^{-2}] = 1/(n-k-2)(n-k-4)$. So taking expectations and collecting terms with powers of a:

$$\begin{aligned}\frac{\Delta_{\hat{\sigma}}}{\sigma^4} \leq \Delta_U(a) \quad = \quad & \left(\frac{\nu(\nu+2)(\nu+4)(\nu+6)}{\nu^4(n-k-2)(n-k-4)}\right)a^4 - \left(\frac{4\nu(\nu+2)(\nu+4)}{\nu^3(n-k-2)}\right)a^3 \\ & + \left(\frac{2\nu(\nu+2)}{\nu^2(n-k-2)} + \frac{4\nu(\nu+2)}{\nu^2}\right)a^2 - 4na\end{aligned}$$

Note that if Δ_U (which does not involve σ) is less than 0, then $\Delta_{\hat{\sigma}} < 0$, which is what we wish to prove.

Finally, note that $\Delta_U(a) = 0$ may be written as $c_4a^4+c_3a^3+c_2a^2+c_1a = 0$, with $c_4 > 0, c_3 < 0, c_2 > 0, c_1 < 0$. The proof then follows exactly as the proof of Theorem 1 in Section 3.1. Since $\Delta_U(a)$ must be negative between its two real roots 0 and r, then $\Delta < 0$ for $0 < a < r$. $\square$

4 Case II: Data Following the Functional Noise Model

Now we will consider functional data following model (2). Again, we assume the response is measured at n discrete points in $[0, T]$, but here we assume a dependence among errors measured at different points.

As with Case I, we assume our linear smoothing matrix $\mathbf{S}$ is symmetric and idempotent, with $r(\mathbf{S}) = tr(\mathbf{S}) = k$. Recall that according to the functional noise model for the data, $\hat{\boldsymbol{\theta}} \sim N(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ where the covariance matrix $\boldsymbol{\Sigma}$ corresponds to a stationary continuous-time process. In this section, we will assume a Gaussian error process whose covariance structure allows for the possibility of dependence at different measurement points.

As with model (1), under model (2), when $\boldsymbol{\theta}$ lies in the linear subspace that $\mathbf{S}$ projects onto, $\frac{T}{n}\hat{\boldsymbol{\theta}}'\mathbf{S}\hat{\boldsymbol{\theta}}$ dominates $\frac{T}{n}\hat{\boldsymbol{\theta}}'\hat{\boldsymbol{\theta}}$ in estimating $\frac{T}{n}\boldsymbol{\theta}'\boldsymbol{\theta}$, as can be seen with a straightforward comparison of MSEs. So let us consider functional data following model (2), without assuming $\boldsymbol{\theta}$ lies in the subspace projected onto by $\mathbf{S}$.

In Section 3, we obtained an exact upper bound for the difference in risks

between the smoothed-data estimator and observed-data estimator for a fixed value of n. In this section we will show that an asymptotic (large n) upper bound exists for this difference of risks. The upper bound is asymptotic here in the sense that the bound is valid for sufficiently large n, not necessarily that the expression for the bound converges to a meaningful limiting expression for infinite n.

Note that as the number of measurement points n grows (within a fixed domain), assuming a certain dependence across measurement points (e.g., a correlation structure like that of a stationary Ornstein-Uhlenbeck process), the observed data $\mathbf{y}_1, \ldots, \mathbf{y}_N$ closely resemble the pure observed functions $y_1(t), \ldots, y_N(t)$ on $[0, T]$. Therefore this result will be most appropriate for situations with a large number of measurements taken on a functional process, so that the observed data vector is "nearly" a pure function.

We consider the same James-Stein estimator of $\boldsymbol{\theta}$ as in Section 3, namely $\hat{\boldsymbol{\theta}}^{(JS)}$ given by (6)—in practice, again, we use the positive-part estimator $\hat{\boldsymbol{\theta}}_+^{(JS)}$ given by (7)—and the same James-Stein dissimilarity estimator.

Theorem 3 *Let $\hat{\boldsymbol{\theta}}$ have known, positive definite covariance matrix $\boldsymbol{\Sigma}$ corresponding to a Gaussian error process. If $n - k > 4$, then there exists a positive real r such that for $0 < a < r$, and for sufficiently large n, the risk difference is negative and the smoothed-data dissimilarity estimator $\hat{d}_{ij}^{(JS)}$ has smaller risk than $\hat{d}_{ij}$.*

Proof of Theorem 3: Recall from Section 3 the notation: $\hat{q}_1 = \hat{\boldsymbol{\theta}}'(\mathbf{I} - \mathbf{S})\hat{\boldsymbol{\theta}}$, $\hat{q}_2 = \hat{\boldsymbol{\theta}}'\mathbf{S}\hat{\boldsymbol{\theta}}$, $q_1 = \boldsymbol{\theta}'(\mathbf{I} - \mathbf{S})\boldsymbol{\theta}$, $q_2 = \boldsymbol{\theta}'\mathbf{S}\boldsymbol{\theta}$.

Unlike in Section 3 when we assumed an independent error structure,

under the functional noise model, $\hat{q}_1$ and $\hat{q}_2$ do not have noncentral χ^2 distributions, nor are they independent in general.

In this same way as in Section 3, we may write Δ as:

$$\begin{aligned}\Delta &= E\Bigg[-4a[\hat{q}_1+\hat{q}_2]+4a^2+\left(\frac{a^2}{\hat{q}_1}\right)^2-\frac{4a^3}{\hat{q}_1}+4a(q_1+q_2)\\ &\quad+\frac{2a^2}{\hat{q}_1}(\hat{q}_1+\hat{q}_2-q_1-q_2)\Bigg].\end{aligned}$$

Since $E[\hat{q}_1+\hat{q}_2]=q_1+tr[(\mathbf{I}-\mathbf{S})\mathbf{\Sigma}]+q_2+tr[\mathbf{S}\mathbf{\Sigma}]=q_1+q_2+tr(\mathbf{\Sigma})$,

$$\Delta=-4atr(\mathbf{\Sigma})+4a^2+E\left[\left(\frac{a^2}{\hat{q}_1}\right)^2-\frac{4a^3}{\hat{q}_1}+\frac{2a^2}{\hat{q}_1}(\hat{q}_1+\hat{q}_2-q_1-q_2)\right].$$

Consider

$$E\left[\frac{\hat{q}_2}{\hat{q}_1}\right]=E\left[\frac{\hat{\boldsymbol{\theta}}'\mathbf{S}\hat{\boldsymbol{\theta}}}{\hat{\boldsymbol{\theta}}'(\mathbf{I}-\mathbf{S})\hat{\boldsymbol{\theta}}}\right].$$

Lieberman (1994) gives a Laplace approximation for $E[(\mathbf{x}'\mathbf{F}\mathbf{x}/\mathbf{x}'\mathbf{G}\mathbf{x})^k]$, $k\geq 1$, where $\mathbf{F}$ is symmetric and $\mathbf{G}$ positive definite:

$$E\left[\left(\frac{\mathbf{x}'\mathbf{F}\mathbf{x}}{\mathbf{x}'\mathbf{G}\mathbf{x}}\right)^k\right]\approx\frac{E[(\mathbf{x}'\mathbf{F}\mathbf{x})^k]}{[E(\mathbf{x}'\mathbf{G}\mathbf{x})]^k}.$$

In our case, $(\mathbf{I}-\mathbf{S})$ is merely positive semidefinite, but with a simple regularity condition, the result will hold (see Appendix A). It can be shown that in this situation the error of the Laplace approximation is $O(1)$ as $n\to\infty$ (see Appendix A).

Hence for large n

$$E\left[\frac{\hat{q}_2}{\hat{q}_1}\right]=E\left[\frac{\hat{\boldsymbol{\theta}}'\mathbf{S}\hat{\boldsymbol{\theta}}}{\hat{\boldsymbol{\theta}}'(\mathbf{I}-\mathbf{S})\hat{\boldsymbol{\theta}}}\right]\leq\frac{E(\hat{\boldsymbol{\theta}}'\mathbf{S}\hat{\boldsymbol{\theta}})}{E(\hat{\boldsymbol{\theta}}'(\mathbf{I}-\mathbf{S})\hat{\boldsymbol{\theta}})}+c=\frac{q_2+tr(\mathbf{S}\mathbf{\Sigma})}{q_1+tr[(\mathbf{I}-\mathbf{S})\mathbf{\Sigma}]}+c$$

for some constant c.

Hence we have a large n approximation, or asymptotic expression, for Δ:

$$\Delta\leq-4atr(\mathbf{\Sigma})+4a^2+E\left[\frac{a^4}{\hat{q}_1^2}+2a^2\left(1+c+\frac{q_2+tr(\mathbf{S}\mathbf{\Sigma})}{q_1+tr[(\mathbf{I}-\mathbf{S})\mathbf{\Sigma}]}-\frac{2a+q_1+q_2}{\hat{q}_1}\right)\right].$$

Since by Jensen's Inequality, $E(1/\hat{q}_1) \geq 1/E(\hat{q}_1)$, we can bound the last term above:

$$\begin{aligned} & E\left[2a^2\left(1+c+\frac{q_2+tr(\mathbf{S\Sigma})}{q_1+tr[(\mathbf{I}-\mathbf{S})\mathbf{\Sigma}]}-\frac{2a+q_1+q_2}{\hat{q}_1}\right)\right] \\ \leq & \ 2a^2\left(1+c+\frac{q_2+tr(\mathbf{S\Sigma})}{q_1+tr[(\mathbf{I}-\mathbf{S})\mathbf{\Sigma}]}-\frac{q_1+q_2+2a}{q_1+tr[(\mathbf{I}-\mathbf{S})\mathbf{\Sigma}]}\right) \\ = & \ 2a^2\left(c+\frac{tr(\mathbf{\Sigma})-2a}{q_1+tr[(\mathbf{I}-\mathbf{S})\mathbf{\Sigma}]}\right) \\ \leq & \ 2a^2\left(c+\frac{tr(\mathbf{\Sigma})-2a}{tr[(\mathbf{I}-\mathbf{S})\mathbf{\Sigma}]}\right). \end{aligned}$$

Assume $c > 0$ (if $c \leq 0$ then it can be ignored without consequence). Then we have the asymptotic upper bound for Δ:

$$4a^2 - 4atr(\mathbf{\Sigma}) + E\left[\frac{a^4}{\hat{q}_1^2}\right] + 2a^2\left(c+\frac{tr(\mathbf{\Sigma})-2a}{tr[(\mathbf{I}-\mathbf{S})\mathbf{\Sigma}]}\right).$$

Using standard eigenvalue properties and distribution theory for quadratic forms (detailed in Appendix B), we have the following asymptotic upper bound for Δ:

$$\Delta_U^* = M_2 a^4 - \frac{4}{tr[(\mathbf{I}-\mathbf{S})\mathbf{\Sigma}]}a^3 + \left(\frac{2tr(\mathbf{\Sigma})}{tr[(\mathbf{I}-\mathbf{S})\mathbf{\Sigma}]}+4+2c\right)a^2 - 4tr(\mathbf{\Sigma})a,$$

where $M_2 = E[(\sum_{i=1}^n e_i\chi_1^2)^{-2}]$, the second inverse moment of a linear combination of independent χ_1^2 variates, with $e_1, \ldots, e_n$ the eigenvalues of $(\mathbf{I}-\mathbf{S})\mathbf{\Sigma}$. We see that if $n-k > 4$, then M_2 exists and is positive since

$$\frac{1}{w_{\max}}E\left[\left(\frac{1}{\chi^2_{n-k}}\right)^2\right] \leq M_2 \leq \frac{1}{w_{\min}}E\left[\left(\frac{1}{\chi^2_{n-k}}\right)^2\right],$$

where $w_{\min}$ is the smallest and $w_{\max}$ the largest of the $n-k$ nonzero eigenvalues of $(\mathbf{I}-\mathbf{S})\mathbf{\Sigma}$.

As was the case in the discrete noise situation, $\Delta_U^*(a) = 0$ is a fourth-degree equation with one nonzero real root r.

Since $\mathbf{\Sigma}$ is positive definite, $tr(\mathbf{\Sigma})$ is positive. Since $M_2 > 0$, we may write $\Delta_U^*(a) = 0$ as $c_4a^4 + c_3a^3 + c_2a^2 + c_1a = 0$, where $c_4 > 0, c_3 < 0, c_2 > 0, c_1 < 0$. The proof then follows exactly from the proof of Theorem 1 in Section 3.1. Since $\Delta_U^*(a)$ must be negative between its two real roots 0 and r, then $\Delta < 0$ for $0 < a < r$ for sufficiently large n. $\square$

One can easily verify that two of the roots are imaginary and can determine the nontrivial real root in terms of M_2, $tr[(\mathbf{I} - \mathbf{S})\mathbf{\Sigma}]$, $tr(\mathbf{\Sigma})$, and c. Though the genuine asymptotic upper bound cannot be calculated since the asymptotic Laplace approximation error c is unknown, its existence is guaranteed. Furthermore, for large n, empirical evidence indicates that the Laplace approximation is quite good, and c is likely to be small. In practice, one may calculate an approximation to the asymptotic upper bound by letting $c = 0$. The resulting fourth-degree function of a, while not guaranteed to be an upper bound for Δ, will probably be close to the true upper bound and at least could be useful in guiding a proper choice of a in the James-Stein estimator.

As an example, let us consider a situation in which we observe functional data $\mathbf{y}_1, \ldots, \mathbf{y}_N$ measured at 30 equally spaced points $t_1, \ldots, t_{30}$, one unit apart. Here, let us assume the observations are discretized versions of functions $y_1(t), \ldots, y_N(t)$ that contain (possibly different) signal functions, plus a noise function arising from an Ornstein-Uhlenbeck (O-U) process. We can then calculate the covariance matrix of each $\mathbf{y}_i$ and the covariance matrix $\mathbf{\Sigma}$ of $\boldsymbol{\theta}$. Under the O-U model, $tr(\mathbf{\Sigma}) = \frac{\sigma^2}{2\beta}n$ always, since each diagonal element

of $\mathbf{\Sigma}$ is $\frac{\sigma^2}{2\beta}$.

For example, suppose the O-U process has $\sigma^2 = 2$ and $\beta = 1$. Then in this example, $tr(\mathbf{\Sigma}) = 30$. Suppose we choose $\mathbf{S}$ be the smoothing matrix corresponding to a B-spline basis smoother with 6 knots, dispersed evenly within the data. Then we can easily calculate the eigenvalues of $(\mathbf{I} - \mathbf{S})\mathbf{\Sigma}$, which are $e_1, \ldots, e_n$, and via numerical or Monte Carlo integration, we find that $M_2 = 0.00758$ in this case.

Substituting these values into $\Delta_U^*(a)$, we see, in Figure 4, the approximation to the asymptotic upper bound plotted as a function of a. Also plotted is a simulated true Δ for a variety of values (n-vectors) of $\boldsymbol{\theta}$: $0 \times \mathbf{1}$, $(0, 1, 0, 1, 0, \ldots, 0, 1)'$, and $(-1, 0, 1, -1, 0, 1, \ldots, -1, 0, 1)'$, where $\mathbf{1}$ is a n-vector of ones. It should be noted that when $\boldsymbol{\theta}$ is the zero vector, it lies in the subspace projected onto by $\mathbf{S}$, since $\mathbf{S}\boldsymbol{\theta} = \boldsymbol{\theta}$ in that case. The other two values of $\boldsymbol{\theta}$ shown in this plot do not lie in the subspace. In this example, choosing the a that minimizes the approximate upper bound ensures that $\hat{d}_{ij}^{(JS)}$ has smaller risk than $\hat{d}_{ij}$.

We now extend the asymptotic result to the case of $\boldsymbol{\theta}$ having covariance matrix of the form $\mathbf{V} = \sigma^2\mathbf{\Sigma}$, where σ^2 is unknown and $\mathbf{\Sigma}$ is a known symmetric, positive definite matrix. This encompasses the functional noise model (2) in which the errors follow an Ornstein-Uhlenbeck process with unknown σ^2 and known β. (Of course, this also includes the discrete noise model (1), in which $\mathbf{V} = \sigma^2\mathbf{I}$, but this case was dealt with in Section 3.2, in which an exact domination result was shown.)

Theorem 4 *Suppose the same conditions as Theorem 2, except generalize the covariance matrix of* $\boldsymbol{\theta}$ *to be* $\sigma^2\mathbf{\Sigma}$, σ^2 *unknown and* $\mathbf{\Sigma}$ *known, symmetric,*

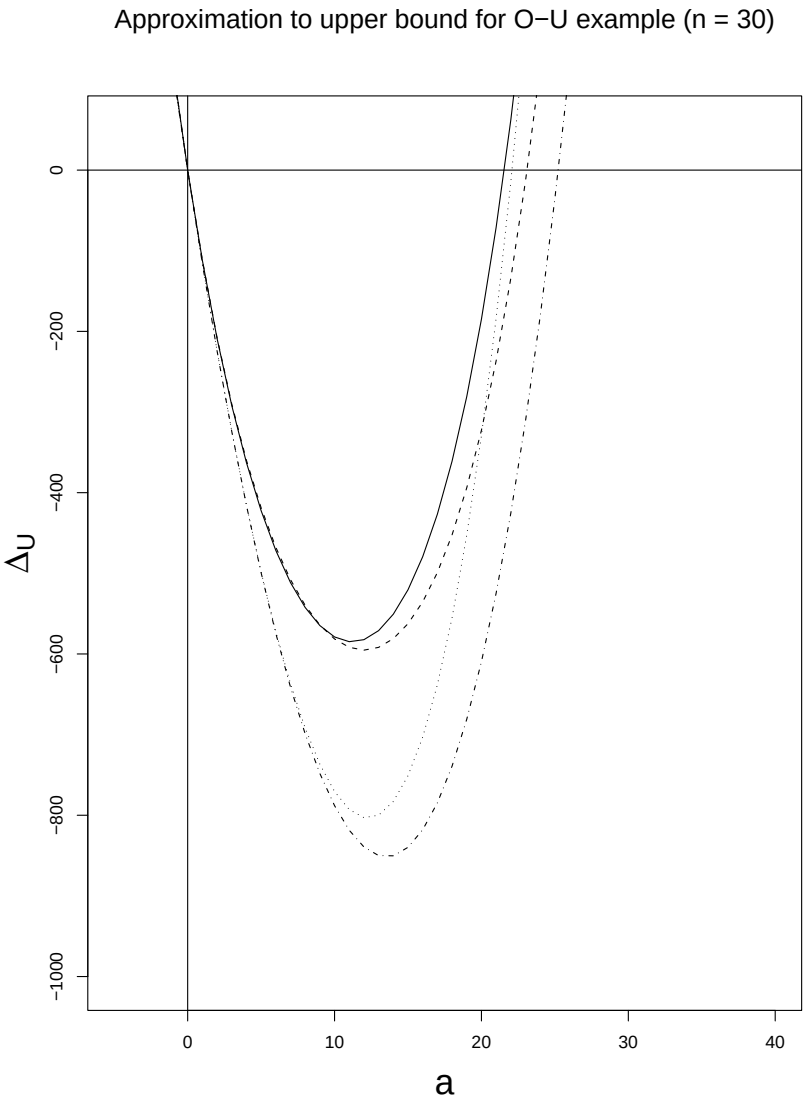

Figure 4: Plot of approximation to asymptotic upper bound, and simulated Δ's, for Ornstein-Uhlenbeck-type data ($n = 30$).

Solid line: plot of approximate Δ_U^* against a for above O-U process. Dashed line: simulated Δ, $\boldsymbol{\theta} = 0 \times \mathbf{1}$. Dotted line: simulated Δ, $\boldsymbol{\theta} = (0, 1, 0, 1, 0, \ldots, 0, 1)'$. Dot-dashed line: simulated Δ, $\boldsymbol{\theta} = (-1, 0, 1, -1, 0, 1, \ldots, -1, 0, 1)'$.

and positive definite. If $n-k>4$, then there exists a positive real r such that for $0<a<r$, and for sufficiently large n, the risk difference is negative and the smoothed-data dissimilarity estimator $\hat{d}^{(JS)}_{ij,\hat{\sigma}}$ has smaller risk than $\hat{d}_{ij}$.

Since much of the proof of Theorem 4 repeats material found in the proofs of Theorem 2 and Theorem 3, we give it in Appendix C.

5 Simulation Study

While the previous theorems guarantee (for a variety of situations) that the James-Stein smoothed dissimilarity estimator has smaller risk than the usual estimator, it may be instructive to study empirically the magnitude of the risk improvement. We can also determine how much the risk improvement is affected if certain assumptions are not met (for example, if the error process is not normal). In this section we summarize a simulation study to help answer these questions. Consider the four signal curves:

$$\begin{aligned}
\mu_1(t) &= -\sin(10t)\ln(t+0.5), t\in[0,1] \\
\mu_2(t) &= \cos(10t)\ln(t+0.5), t\in[0,1] \\
\mu_3(t) &= 0.25\sin(10t)\sqrt{5t^{1/2}+0.5}, t\in[0,1] \\
\mu_4(t) &= 0.25+\cos(10t)\ln(t+0.5)/\sqrt{t+0.5}, t\in[0,1]
\end{aligned}$$

From these signal curves we generate noisy functional data (observed at $n=30$ points in $[0,1]$), by adding one of three types of random noise: independent $N(0,\sigma^2)$ with varying σ^2; Ornstein-Uhlenbeck with varying σ^2 and $\beta=1$; or heavy-tailed noise (independent t with varying degrees of freedom). (Note

that the values of σ^2 are not directly comparable between the independent and Ornstein-Uhlenbeck models.) We smooth the resulting $\hat{\boldsymbol{\theta}}$ values with a B-spline-based smoother having 6 knots interspersed evenly though the data (implying $k = 10$). The James-Stein adjustment is made with $a = 20$ (based on $n = 30, k = 10$) for the independent error data and appropriate values of a for the respective choices of $\boldsymbol{\Sigma}$ for the dependent error data.

These four curves define six pairwise dissimilarities d_{ij}, $(i, j) \in \{1, 2, 3, 4\}$, $i < j$, which may be estimated by $\hat{d}_{ij}$ or $\hat{d}_{ij}^{(JS)}$. Define $\hat{\Delta} = \overline{MSE}(\hat{d}_{ij}^{(JS)}) - \overline{MSE}(\hat{d}_{ij})$, where $MSE(\hat{d}_{ij}) = (1/6)\sum_{i<j}(\hat{d}_{ij} - d_{ij})^2$. So $\hat{\Delta}$ is the difference of the empirical risks (averaged over $50,000$ iterations) of $\hat{d}_{ij}$ and $\hat{d}_{ij}^{(JS)}$. Shown in Table 2 are values of $\hat{\Delta}$ (with Monte Carlo standard deviations for $\hat{\Delta}$ in parentheses) for the various noise distributions and magnitudes of error variability. Negative values of $\hat{\Delta}$ in the table indicate the James-Stein estimator has smaller risk. The values in braces indicating the percentage risk improvement from using the James-Stein estimator are more meaningful than the raw $\hat{\Delta}$ values, which vary greatly depending on how noisy the data are.

The results show that the James-Stein estimator has, on average, smaller empirical risk in every case examined except the t (2 df) case. The amount of risk improvement varies according to the magnitude and type of noise, mostly ranging from around 50 to 85 % improvement (an exception being the Ornstein-Uhlenbeck model with $\sigma^2 = 0.5$, in which the noise in the data is minimal). It appears that, in the normal cases, the improvement increases as the error variability grows. The risk improvement appears to hold for the t errors, only breaking down for the extremely heavy-tailed case of 2 degrees

Table 2: Empirical risk differences between James-Stein dissimilarity estimator and observed-data dissimilarity estimator.

Independent Normal Errors			
$\sigma^2 = 0.5$	$\sigma^2 = 1$	$\sigma^2 = 2$	$\sigma^2 = 4$
-2.74 (0.73)	-2.42 (0.63)	-2.40 (0.61)	-2.42 (0.61)
{56.4%}	{65.1%}	{66.7%}	{66.8%}
Ornstein-Uhlenbeck Errors			
$\sigma^2 = 0.5$	$\sigma^2 = 1$	$\sigma^2 = 2$	$\sigma^2 = 4$
-1.6×10^{-5} (0.0001)	-0.002 (0.002)	-0.05 (0.02)	-0.87 (0.30)
{4.8%}	{58.3%}	{81.3%}	{85.8%}
Independent t Errors			
d.f. = 20	d.f. = 10	d.f. = 5	d.f. = 2
-0.23 (0.08)	-0.29 (0.12)	-0.55 (0.36)	11131 (2×10^6)
{87.5%}	{87.1%}	{82.8%}	{-1.8%}

NOTE: Monte Carlo s.d. is in parentheses; percentage risk improvement for James-Stein estimator is in braces.

of freedom.

6 Discussion

This paper has addressed the problem of estimating dissimilarities for functional observations. The dissimilarities among objects in a data set are at the heart of distance-based statistical methods such as cluster analysis, multidimensional scaling and statistical matching.

We have proposed a model for the functional observations and hence for the dissimilarities among them. When the data are smoothed using a basis function method (such as regression splines, for example), we have shown that a James-Stein shrinkage dissimilarity estimator dominates the observed-data estimator under an independent error model. With dependent errors, an asymptotic (for n large within a fixed domain) domination result was given. (Note that the asymptotic situation of the theorem corresponds to data that are nearly pure functions, measured nearly continuously across some domain.)

A simulation study has indicated the magnitude of the risk improvement and suggested that the results hold for moderately heavy-tailed non-normal errors. The shrinkage estimator is a novel way to unite linear smoothers and Stein estimation to derive a useful, data-informed smoothing method. It is hoped that these results contribute to resolving the increasing need for methods of analyzing large functional data sets.

A Notes about Laplace Approximation

Recall the Laplace approximation given by Lieberman (1994) for the expectation of a ratio of quadratic forms:

$$E\left(\frac{\mathbf{x}'\mathbf{F}\mathbf{x}}{\mathbf{x}'\mathbf{G}\mathbf{x}}\right) \approx \frac{E(\mathbf{x}'\mathbf{F}\mathbf{x})}{E(\mathbf{x}'\mathbf{G}\mathbf{x})}.$$

Lieberman (1994) denotes the joint moment generating function of $\mathbf{x}'\mathbf{F}\mathbf{x}$ and $\mathbf{x}'\mathbf{G}\mathbf{x}$ by

$$M(\omega_1, \omega_2) = E[\exp(\omega_1 \mathbf{x}'\mathbf{F}\mathbf{x} + \omega_2 \mathbf{x}'\mathbf{G}\mathbf{x})].$$

and assumes a positive definite $\mathbf{G}$. In that case, $E(\mathbf{x}'\mathbf{G}\mathbf{x}) > 0$. Lieberman uses the positive definiteness of $\mathbf{G}$ to show that the derivative of the cumulant generating function of $\mathbf{x}'\mathbf{G}\mathbf{x}$ is greater than zero. That is,

$$\begin{aligned}
\frac{d}{d\omega_2}\log M(0,\omega_2) &= \frac{d}{d\omega_2}M(0,\omega_2)\left[M(0,\omega_2)\right]^{-1} \\
&= \int(\mathbf{x}'\mathbf{G}\mathbf{x})\exp\{\omega_2\mathbf{x}'\mathbf{G}\mathbf{x}\}f(\mathbf{x})\,d\mathbf{x}\left[\int\exp\{\omega_2\mathbf{x}'\mathbf{G}\mathbf{x}\}f(\mathbf{x})\,d\mathbf{x}\right]^{-1} \\
&> 0.
\end{aligned}$$

The positive derivative ensures the maximum of $\log M(0,\omega_2)$ is attained at the boundary point (where $\omega_2 = 0$).

For positive semidefinite $\mathbf{G}$, we need the additional regularity condition that $P(\mathbf{x}'\mathbf{G}\mathbf{x} > 0) > 0$, i.e., the support of $\mathbf{x}'\mathbf{G}\mathbf{x}$ is not degenerate at zero. This will ensure that $E(\mathbf{x}'\mathbf{G}\mathbf{x}) > 0$, i.e., that $\frac{d}{d\omega_2}M(0,\omega_2) > 0$. Therefore both of the integrals in the above expression are positive and the Laplace approximation will hold.

In stating the order of the error of the Laplace approximation, we refer to Lieberman's (1994) sufficient conditions for the error to be $O(n^{-1})$. We assume (reasonably, thanks to the smoothness properties of $\mathbf{S}$ and the fact that $(T/n)\boldsymbol{\theta}'\mathbf{S}\boldsymbol{\theta}$ approximates an L_2 distance) that as $n \to \infty$,

$$\frac{T}{n}\boldsymbol{\theta}'\mathbf{S}\boldsymbol{\theta} \to c_1$$

where c_1 is some constant.

Note that $tr(\mathbf{S}\boldsymbol{\Sigma}) \leq tr(\mathbf{S}\boldsymbol{\Sigma}) + tr((\mathbf{I}-\mathbf{S})\boldsymbol{\Sigma}) = tr(\boldsymbol{\Sigma}) = \sigma^2 n/\beta$ under the Ornstein-Uhlenbeck model. Hence as $n \to \infty$,

$$\begin{aligned}
\frac{E[\hat{\boldsymbol{\theta}}'\mathbf{S}\hat{\boldsymbol{\theta}}]}{n} &= \frac{\boldsymbol{\theta}'\mathbf{S}\boldsymbol{\theta}}{n} + \frac{tr(\mathbf{S}\boldsymbol{\Sigma})}{n} \\
&\leq \frac{\boldsymbol{\theta}'\mathbf{S}\boldsymbol{\theta}}{n} + \sigma^2/\beta \to c_1/T + \sigma^2/\beta = c_2
\end{aligned}$$

where c_2 is a constant. Thus $E[\hat{\boldsymbol{\theta}}'\mathbf{S}\hat{\boldsymbol{\theta}}] = O(n)$. Following similarly are the orders of higher moments, e.g., $E[(\hat{\boldsymbol{\theta}}'\mathbf{S}\hat{\boldsymbol{\theta}})^k] = O(n^k), k \geq 1$. On the other hand, since $E[\hat{\boldsymbol{\theta}}'\mathbf{S}\hat{\boldsymbol{\theta}}] = O(n)$ and similarly $E[\hat{\boldsymbol{\theta}}'(\mathbf{I}-\mathbf{S})\hat{\boldsymbol{\theta}}] = O(n)$, then $cov[\hat{\boldsymbol{\theta}}'\mathbf{S}\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}}'(\mathbf{I}-\mathbf{S})\hat{\boldsymbol{\theta}}] = O(n^2)$, rather than $O(n)$ as Lieberman's third condition requires. This implies the Laplace approximation has an error of $O(1)$ rather than $O(n^{-1})$ had all three of Lieberman's conditions been met.

B Additional Details of Proof of Theorem 3

Denote the eigenvalues of $\boldsymbol{\Sigma}^{1/2}(\mathbf{I}-\mathbf{S})\boldsymbol{\Sigma}^{1/2}$ by $e_1, \ldots, e_n$. Since $(\mathbf{I}-\mathbf{S})$ is positive semidefinite, it is clear that $e_1, \ldots, e_n$ are all nonnegative. Note that $e_1, \ldots, e_n$ are also the eigenvalues of $(\mathbf{I}-\mathbf{S})\boldsymbol{\Sigma}$, since $\boldsymbol{\Sigma}^{1/2}(\mathbf{I}-\mathbf{S})\boldsymbol{\Sigma}^{1/2} = \boldsymbol{\Sigma}^{1/2}(\mathbf{I}-\mathbf{S})\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1/2}$ and $(\mathbf{I}-\mathbf{S})\boldsymbol{\Sigma}$ are similar matrices. Note that since $r[\boldsymbol{\Sigma}^{1/2}(\mathbf{I}-\mathbf{S})\boldsymbol{\Sigma}^{1/2}] = r(\mathbf{I}-\mathbf{S}) = n-k$, $\boldsymbol{\Sigma}^{1/2}(\mathbf{I}-\mathbf{S})\boldsymbol{\Sigma}^{1/2}$ has $n-k$ nonzero eigenvalues, and so does $(\mathbf{I}-\mathbf{S})\boldsymbol{\Sigma}$.

It is well known (see, e.g., Baldessari 1967; Tan 1977) that if $\mathbf{y} \sim N(\boldsymbol{\mu}, \mathbf{V})$ for positive definite, nonsingular $\mathbf{V}$, then for symmetric $\mathbf{A}$, $\mathbf{y}'\mathbf{A}\mathbf{y}$ is distributed as a linear combination of independent noncentral χ^2 random variables, the coefficients of which are the eigenvalues of $\mathbf{AV}$.

Specifically, since $\hat{q}_1 = \hat{\boldsymbol{\theta}}'(\mathbf{I}-\mathbf{S})\hat{\boldsymbol{\theta}}$, and $(\mathbf{I}-\mathbf{S})\boldsymbol{\Sigma}$ has eigenvalues $e_1, \ldots, e_n$, then

$$\hat{q}_1 \sim \sum_{i=1}^{n} e_i \chi_1^2(\delta_i^2)$$

for some noncentrality parameter $\delta_i^2 \geq 0$ that is zero when $\boldsymbol{\theta} = \mathbf{0}$.

Under $\boldsymbol{\theta} = \mathbf{0}$, then, $\hat{q}_1 \sim \sum_{i=1}^{n} e_i \chi_1^2$, and since a noncentral χ_1^2 random

variable stochastically dominates a central χ^2_1, for any $\delta_i^2 \geq 0$,

$$P[\chi^2_1(\delta_i^2) \geq x] \geq P[\chi^2_1 \geq x] \ \ \forall\, x > 0.$$

We know that for each $\delta_i^2 \geq 0$, $i = 1, \ldots, n$,

$$P[\chi^2_1(\delta_i^2) \geq x] \geq P[\chi^2_1 \geq x] \ \ \forall\, x > 0.$$

Since $e_1, \ldots, e_n \geq 0$, letting $z_i = e_i x$, $i = 1, \ldots, n$,

$$P[e_i\chi^2_1(\delta_i^2) \geq z_i] \geq P[e_i\chi^2_1 \geq z_i] \ \ \forall\, z_i > 0.$$

(If $e_i = 0$, the inequality trivially holds.) Now, since $f(x_1, \ldots, x_n) = x_1 + \ldots + x_n$ is an increasing function, and since the n noncentral chi-squares are independent and the n central chi-squares are independent, we apply a result given in Ross (1996, p. 410, ex. 9.2(A)) to conclude:

$$P[e_1\chi^2_1(\delta_1^2) + \cdots + e_n\chi^2_1(\delta_n^2) \geq x] \geq P[e_1\chi^2_1 + \cdots + e_n\chi^2_1 \geq x] \ \ \forall \ \ x > 0. \ \ \square$$

Hence for all $x > 0$, $P_{\boldsymbol{\theta} \neq \mathbf{0}}[\hat{q}_1 \geq x] \geq P_{\mathbf{0}}[\hat{q}_1 \geq x]$, i.e., the distribution of $\hat{q}_1$ with $\boldsymbol{\theta} \neq \mathbf{0}$ stochastically dominates the distribution of $\hat{q}_1$ with $\boldsymbol{\theta} = \mathbf{0}$. Then

$$E_{\boldsymbol{\theta} \neq \mathbf{0}}[(\hat{q}_1)^{-m}] \leq E_{\mathbf{0}}[(\hat{q}_1)^{-m}]$$

for $m = 1, 2, \ldots$.

C Proof of Theorem 4

Assume $\hat{\boldsymbol{\theta}} \sim N(\boldsymbol{\theta}, \mathbf{V})$, with $\mathbf{V} = \sigma^2\boldsymbol{\Sigma}$. Suppose, as in Section 3.2, that there exists a random variable S^2, independent of $\hat{\boldsymbol{\theta}}$, such that $S^2/\sigma^2 \sim \chi^2_\nu$. Then let $\hat{\sigma}^2 = S^2/\nu$.

As shown in Section 3.2, we may write

$$\Delta_{\hat{\sigma}} = E\left[-2\left(2a\hat{\sigma}^2 - \frac{a^2\hat{\sigma}^4}{\hat{q}_1}\right)(\hat{\boldsymbol{\theta}}'\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}'\boldsymbol{\theta}) + \left(2a\hat{\sigma}^2 - \frac{a^2\hat{\sigma}^4}{\hat{q}_1}\right)^2\right].$$

Since $\hat{\sigma}$ and $\hat{\boldsymbol{\theta}}$ are independent, recall

$$E[-4a\hat{\sigma}^2(\hat{\boldsymbol{\theta}}'\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}'\boldsymbol{\theta})] = E[-4a\hat{\sigma}^2]\sigma^2 tr(\boldsymbol{\Sigma})$$

and recall from Section 3.2,

$$\begin{aligned} & E\left[\frac{2a^2\hat{\sigma}^4}{\hat{q}_1}(\hat{\boldsymbol{\theta}}'\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}'\boldsymbol{\theta})\right] \\ = \;& E[2a^2\hat{\sigma}^4]E\left[\frac{\hat{q}_2}{\hat{q}_1}\right] + E\left[\frac{2a^2\hat{\sigma}^4}{\hat{q}_1}(\hat{q}_1 - q_1 - q_2)\right]. \end{aligned} \tag{16}$$

We use the (large n) approximation to $E[\hat{q}_2/\hat{q}_1]$ obtained in Section 4 to obtain an asymptotic upper bound for (16):

$$E[2a^2\hat{\sigma}^4]\left(\frac{q_2 + \sigma^2 tr(\mathbf{S}\boldsymbol{\Sigma})}{q_1 + \sigma^2 tr[(\mathbf{I}-\mathbf{S})\boldsymbol{\Sigma}]} + c\right) + E\left[\frac{2a^2\hat{\sigma}^4}{\hat{q}_1}(\hat{q}_1 - q_1 - q_2)\right]. \tag{17}$$

Now, since $a^2\hat{\sigma}^4/\hat{q}_1$ is decreasing in $\hat{q}_1$ and $\hat{q}_1 - q_1 - q_2$ is increasing in $\hat{q}_1$, these quantities have covariance ≤ 0. Hence

$$\begin{aligned} (17) \;\leq\; & E[2a^2\hat{\sigma}^4]\left(\frac{q_2 + \sigma^2 tr(\mathbf{S}\boldsymbol{\Sigma})}{q_1 + \sigma^2 tr[(\mathbf{I}-\mathbf{S})\boldsymbol{\Sigma}]} + c\right) + E\left[\frac{2a^2\hat{\sigma}^4}{\hat{q}_1}\right]E[\hat{q}_1 - q_1 - q_2] \\ = \;& E[2a^2\hat{\sigma}^4]q_2\left[\frac{1}{E[\hat{q}_1]} - E\left[\frac{1}{\hat{q}_1}\right]\right] + E[2a^2\hat{\sigma}^4]\frac{\sigma^2 tr(\mathbf{S}\boldsymbol{\Sigma})}{q_1 + \sigma^2 tr[(\mathbf{I}-\mathbf{S})\boldsymbol{\Sigma}]} \\ & + E\left[\frac{2a^2\hat{\sigma}^4}{\hat{q}_1}\right]\sigma^2 tr[(\mathbf{I}-\mathbf{S})\boldsymbol{\Sigma}] + E[2a^2\hat{\sigma}^4]c \\ \leq \;& E[2a^2\hat{\sigma}^4]\frac{tr(\mathbf{S}\boldsymbol{\Sigma})}{tr[(\mathbf{I}-\mathbf{S})\boldsymbol{\Sigma}]} + E\left[\frac{2a^2\hat{\sigma}^4}{\hat{q}_1}\right]\sigma^2 tr[(\mathbf{I}-\mathbf{S})\boldsymbol{\Sigma}] + E[2a^2\hat{\sigma}^4]c, \end{aligned}$$

Using the fact that $1/E[\hat{q}_1] - E[1/\hat{q}_1] \leq 0$. Our asymptotic upper bound for

$\Delta_{\hat{\sigma}}$ is:

$$\begin{aligned}\Delta^*_{\hat{\sigma}U} &= E\Bigg[-4a\hat{\sigma}^2\sigma^2 tr(\boldsymbol{\Sigma}) + \frac{2a^2\hat{\sigma}^4}{\hat{q}_1}\sigma^2 tr[(\mathbf{I}-\mathbf{S})\boldsymbol{\Sigma}] + 2a^2\hat{\sigma}^4\left(\frac{tr(\mathbf{S}\boldsymbol{\Sigma})}{tr[(\mathbf{I}-\mathbf{S})\boldsymbol{\Sigma}]}+c\right) \\ &\quad + 4a^2\hat{\sigma}^4 - \frac{4a^3\hat{\sigma}^6}{\hat{q}_1} + \frac{a^4\hat{\sigma}^8}{\hat{q}_1^2}\Bigg].\end{aligned}$$

Recall that $E[a\hat{\sigma}^2] = a\sigma^2$, $E[a^2\hat{\sigma}^4] = a^2\sigma^4\nu(\nu+2)/\nu^2$, $E[a^3\hat{\sigma}^6] = a^3\sigma^6\nu(\nu+2)(\nu+4)/\nu^3$, $E[a^4\hat{\sigma}^8] = a^4\sigma^8\nu(\nu+2)(\nu+4)(\nu+6)/\nu^4$.

Since $\hat{\sigma}^2$ and $\hat{q}_1$ are independent, taking expectations:

$$\begin{aligned}\Delta^*_{\hat{\sigma}U} &= -4a\sigma^4 tr(\boldsymbol{\Sigma}) + 2a^2\sigma^4\left(\frac{\nu(\nu+2)}{\nu^2}\right)E\left[\frac{1}{\hat{q}_1}\right]\sigma^2 tr[(\mathbf{I}-\mathbf{S})\boldsymbol{\Sigma}] \\ &\quad + 2a^2\sigma^4\left(\frac{\nu(\nu+2)}{\nu^2}\right)\left(\frac{tr(\mathbf{S}\boldsymbol{\Sigma})}{tr[(\mathbf{I}-\mathbf{S})\boldsymbol{\Sigma}]}+c\right) + 4a^2\sigma^4\left(\frac{\nu(\nu+2)}{\nu^2}\right) \\ &\quad - 4a^3\sigma^6\left(\frac{\nu(\nu+2)(\nu+4)}{\nu^3}\right)E\left[\frac{1}{\hat{q}_1}\right] + a^4\sigma^8\left(\frac{\nu(\nu+2)(\nu+4)(\nu+6)}{\nu^4}\right)E\left[\frac{1}{\hat{q}_1^2}\right].\end{aligned}$$

As in Section 3.2, define $\hat{q}_1^* = \hat{q}_1/\sigma^2 = (\hat{\boldsymbol{\theta}}/\sigma)'(\mathbf{I}-\mathbf{S})(\hat{\boldsymbol{\theta}}/\sigma)$. Here, since $\hat{\boldsymbol{\theta}} \sim N(\boldsymbol{\theta}, \sigma^2\boldsymbol{\Sigma})$, $(\hat{\boldsymbol{\theta}}/\sigma) \sim N(\boldsymbol{\zeta}, \boldsymbol{\Sigma})$ where $\boldsymbol{\zeta} = \boldsymbol{\theta}/\sigma$. Then the risk difference is a function of $\boldsymbol{\zeta}$, and the distribution of $\hat{q}_1^*$ is a function of $\boldsymbol{\zeta}$. Then (again dividing the inequality through by $\sigma^4 > 0$)

$$\begin{aligned}\frac{\Delta^*_{\hat{\sigma}U}}{\sigma^4} &= -4a tr(\boldsymbol{\Sigma}) + 2a^2\left(\frac{\nu(\nu+2)}{\nu^2}\right)E\left[\frac{1}{\hat{q}_1^*}\right] tr[(\mathbf{I}-\mathbf{S})\boldsymbol{\Sigma}] \\ &\quad + 2a^2\left(\frac{\nu(\nu+2)}{\nu^2}\right)\left(\frac{tr(\mathbf{S}\boldsymbol{\Sigma})}{tr[(\mathbf{I}-\mathbf{S})\boldsymbol{\Sigma}]}+c\right) + 4a^2\left(\frac{\nu(\nu+2)}{\nu^2}\right) \\ &\quad - 4a^3\left(\frac{\nu(\nu+2)(\nu+4)}{\nu^3}\right)E\left[\frac{1}{\hat{q}_1^*}\right] + a^4\left(\frac{\nu(\nu+2)(\nu+4)(\nu+6)}{\nu^4}\right)E\left[\frac{1}{\hat{q}_1^{*2}}\right].\end{aligned}$$

As shown in Section 4, for all $x > 0$, $P_{\boldsymbol{\theta}\neq\mathbf{0}}[\hat{q}_1 \geq x] \geq P_{\mathbf{0}}[\hat{q}_1 \geq x]$. Note that this is equivalent to: For all $x > 0$, $P_{\boldsymbol{\zeta}\neq\mathbf{0}}[\hat{q}_1^* \geq x] \geq P_{\mathbf{0}}[\hat{q}_1^* \geq x]$. Then

$$E_{\boldsymbol{\zeta}\neq\mathbf{0}}[(\hat{q}_1^*)^{-m}] \leq E_{\mathbf{0}}[(\hat{q}_1^*)^{-m}]$$

for $m = 1, 2$.

Let $M_1^* = E_{\mathbf{0}}[(\hat{q}_1^*)^{-1}]$ and $M_2^* = E_{\mathbf{0}}[(\hat{q}_1^*)^{-2}]$. Then, collecting terms with powers of a,

$$\begin{aligned}\frac{\Delta^*_{\hat{\sigma}U}}{\sigma^4} \leq & \left(\frac{\nu(\nu+2)(\nu+4)(\nu+6)}{\nu^4}\right) M_2^* a^4 - 4\left(\frac{\nu(\nu+2)(\nu+4)}{\nu^3}\right) M_1^* a^3 \\ & + \left(\frac{\nu(\nu+2)}{\nu^2}(2M_1^* tr[(\mathbf{I}-\mathbf{S})\mathbf{\Sigma}] + \frac{2tr(\mathbf{S}\mathbf{\Sigma})}{tr[(\mathbf{I}-\mathbf{S})\mathbf{\Sigma}]} + 2c + 4)\right) a^2 - 4tr(\mathbf{\Sigma})a.\end{aligned}$$

Note that if this last expression (which does not involve σ and which we denote simply as $\Delta_U(a)$) is less than 0, then $\Delta^*_{\hat{\sigma}U} < 0$, which is what we wish to prove.

We may repeat the argument from Section 4 in which we showed that when $\boldsymbol{\theta} = \mathbf{0}$, $\hat{q}_1 \sim \sum_{i=1}^n e_i \chi_1^2$, replacing $\hat{q}_1$ with $\hat{q}_1^*$, $\hat{\boldsymbol{\theta}}$ with $\hat{\boldsymbol{\theta}}/\sigma$, and $\boldsymbol{\theta}$ with $\boldsymbol{\zeta}$. Thus we conclude that when $\boldsymbol{\zeta} = \mathbf{0}$, $\hat{q}_1^* \sim \sum_{i=1}^n v_i \chi_1^2$, where $v_1, \ldots, v_n$ are the eigenvalues of $(\mathbf{I}-\mathbf{S})\mathbf{\Sigma}$. Again, if $n - k > 4$, then M_1^* and M_2^* exist and are positive, as was shown in Section 4.

Again, $\Delta_U(a) = 0$ is a fourth-degree equation with one nonzero real root r. Since $\mathbf{\Sigma}$ is positive definite, $tr(\mathbf{\Sigma})$ is positive. Since $M_1^* > 0$ and $M_2^* > 0$, we may write $\Delta_U(a) = 0$ as $c_4 a^4 + c_3 a^3 + c_2 a^2 + c_1 a = 0$, where $c_4 > 0, c_3 < 0, c_2 > 0, c_1 < 0$. The proof follows exactly as the proof of Theorem 3 in Section 4. Since $\Delta_U(a)$ must be negative between its two real roots 0 and r, then $\Delta < 0$ for $0 < a < r$ for sufficiently large n. $\square$

References

[1] Baldessari, B. (1967), "The Distribution of a Quadratic Form of Normal Random Variables," *The Annals of Mathematical Statistics*, 38, 1700–

1704.

[2] Buja, A., Hastie, T., and Tibshirani, R. (1989), "Linear Smoothers and Additive Models," *The Annals of Statistics*, 17, 453–510.

[3] Casella, G., and Hwang, J. T. (1987), "Employing Vague Prior Information in the Construction of Confidence Sets," *Journal of Multivariate Analysis*, 21, 79–104.

[4] Cressie, N. A. C. (1993), *Statistics for Spatial Data*, New York: John Wiley and Sons.

[5] de Boor, C. (1978), *A Practical Guide to Splines*, Redwood City, CA: Addison-Wesley.

[6] Eubank, R. L. (1988), *Spline Smoothing and Nonparametric Regression*, New York: Marcel Dekker Inc.

[7] Everitt, B., Landau, S., and Leese, M. (2001), *Cluster Analysis*, London: Edward Arnold Publishers Ltd.

[8] Falkner, F. (ed.) (1960), *Child Development: An International Method of Study*, Basel: Karger.

[9] Green, E. J., and Strawderman, W. E. (1991), "A James-Stein Type Estimator for Combining Unbiased and Possibly Biased Estimators," *Journal of the American Statistical Association*, 86, 1001–1006.

[10] Johnson, R. A., and Wichern, D. W. (1998), *Applied Multivariate Statistical Analysis*, Upper Saddle River, NJ: Prentice Hall Inc.

[11] Kaufman, L., and Rousseeuw, P. J. (1987), "Clustering by Means of Medoids," in *Statistical Data Analysis Based on the L_1 Norm*, pp. 405–416.

[12] Kaufman, L., and Rousseeuw, P. J. (1990), *Finding Groups in Data: An Introduction to Cluster Analysis*, New York: John Wiley and Sons.

[13] Lehmann, E. L., and Casella, G. (1998), *Theory of Point Estimation*, New York: Springer-Verlag Inc.

[14] Lieberman, O. (1994), "A Laplace Approximation to the Moments of a Ratio of Quadratic Forms," *Biometrika*, 81, 681–690.

[15] Ramsay, J. O. and Dalzell, C. J. (1991), "Some Tools for Functional Data Analysis," *Journal of the Royal Statistical Society, Series B, Methodological*, 53, 539–561.

[16] Ramsay, J. O., and Silverman, B. W. (1997), *Functional Data Analysis*, New York: Springer-Verlag Inc.

[17] Rodgers, W. L. (1988), "Statistical Matching," in *Encyclopedia of Statistical Sciences (9 Vols. Plus Supplement), Volume 8*, pp. 663–664.

[18] Ross, S. M. (1996), *Stochastic Processes*, New York: John Wiley and Sons.

[19] Stein, M. L. (1995), "Locally Lattice Sampling Designs for Isotropic Random Fields," *The Annals of Statistics*, 23, 1991–2012.

[20] Tan, W. Y. (1977), "On the Distribution of Quadratic Forms in Normal Random Variables," *The Canadian Journal of Statistics*, 5, 241–250.

[21] Taylor, J. M. G., Cumberland, W. G., and Sy, J. P. (1994), "A Stochastic Model for Analysis of Longitudinal AIDS Data," *Journal of the American Statistical Association*, 89, 727–736.

[22] Young, F. W., and Hamer, R. M. (1987), *Multidimensional Scaling: History, Theory, and Applications*, Hillsdale, NJ: Lawrence Erlbaum Associates.