Supplementary Material for "The Falling Factorial Basis and Its Statistical Applications"

Yu-Xiang Wang yuxiangw@cs.cmu.edu
Alex Smola alex@smola.org
Ryan J. Tibshirani ryantibs@stat.cmu.edu

This document contains proofs and additional experiments for the paper "The Falling Factorial Basis and Its Statistical Applications". In Section A, we provide proofs to the key technical results in the main paper. In Section B, we give some motivating arguments and additional experiments for the higher order KS test.

A Proofs and technical details

A.1 Proof of Lemma 1 (recursive decomposition)

The falling factorial basis matrix, as defined in (4), (5), can be expressed as $H^{(k)} = [H_1^{(k)}\ H_2^{(k)}]$, where

$$H_1^{(k)} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & x_2 - x_1 & 0 & \cdots & 0 \\ 1 & x_3 - x_1 & (x_3 - x_2)(x_3 - x_1) & \cdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{k+1} - x_1 & (x_{k+1} - x_2)(x_{k+1} - x_1) & \cdots & \prod_{\ell=1}^{k}(x_{k+1} - x_\ell) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n - x_1 & (x_n - x_2)(x_n - x_1) & \cdots & \prod_{\ell=1}^{k}(x_n - x_\ell) \end{bmatrix} \in \mathbb{R}^{n \times (k+1)},$$

and

$$H_2^{(k)} = \begin{bmatrix} 0_{(k+1)\times 1} & 0_{(k+1)\times 1} & \cdots & 0_{(k+1)\times 1} \\ \prod_{\ell=1}^{k}(x_{k+2} - x_{1+\ell}) & 0 & \cdots & 0 \\ \prod_{\ell=1}^{k}(x_{k+3} - x_{1+\ell}) & \prod_{\ell=1}^{k}(x_{k+3} - x_{2+\ell}) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \prod_{\ell=1}^{k}(x_n - x_{1+\ell}) & \prod_{\ell=1}^{k}(x_n - x_{2+\ell}) & \cdots & \prod_{\ell=1}^{k}(x_n - x_{n-k-1+\ell}) \end{bmatrix} \in \mathbb{R}^{n \times (n-k-1)}.$$

Lemma 1 claims that $H^{(0)} = L_n$, the lower triangular matrix of 1s, which can be seen directly by inspection (recalling our convention of defining thee empty product to be 1). The lemma further claims that $H^{(k)}$ can be recursively factorized into the following form:

$$H^{(k)} = H^{(k-1)} \cdot \begin{bmatrix} I_k & 0 \\ 0 & \Delta^{(k)} \end{bmatrix} \cdot \begin{bmatrix} I_k & 0 \\ 0 & L_{n-k} \end{bmatrix}, \tag{A.1}$$

for all $k \geq 1$. We prove the above factorization in this current section. In what follows, we denote the last $n - k - 1$ columns of the product (A.1) by $\tilde{M}^{(k)} \in \mathbb{R}^{n \times (n-k-1)}$, and also write

$$\tilde{M}^{(k)} = \begin{bmatrix} 0_{(k+1)\times(n-k-1)} \\ \tilde{L}^{(k)}, \end{bmatrix},$$

i.e., we use $\tilde{L}^{(k)}$ to denote the lower $(n-k-1)\times(n-k-1)$ submatrix of $\tilde{M}^{(k)}$. To prove the lemma, we show that $\tilde{M}^{(k)}$ is equal to the corresponding block $H_2^{(k)}$, by induction on k. The proof that the first block of $k+1$ columns of the product is equal to $H_1^{(k)}$ follows from the arguments given for the proof of the second block, and therefore we do not explicitly rewrite the proof for this part.

We begin the inductive proof by checking the case $k=1$. Note

$$\begin{aligned}\tilde{M}^{(1)} = \begin{bmatrix} 0_{2\times(n-2)} \\ \tilde{L}^{(1)} \end{bmatrix} &= \begin{bmatrix} 0_{1\times(n-1)} \\ L_{n-1} \end{bmatrix} (\Delta^{(k)})^{-1} \begin{bmatrix} 0_{1\times(n-2)} \\ L_{n-2} \end{bmatrix} \\ &= \begin{bmatrix} 0_{2\times1} & 0_{2\times1} & \cdots & 0_{2\times1} \\ x_3-x_2 & 0 & \cdots & 0 \\ x_4-x_2 & x_4-x_3 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ x_n-x_2 & x_n-x_3 & \cdots & x_n-x_{n-1} \end{bmatrix}.\end{aligned}$$

This gives precisely the last $n-2$ columns of $H^{(1)}$, as defined in (4).

Next we verify that if the statement holds for some $k\geq 1$, then it is true for $k+1$. To avoid confusion, we will use i,j as indices $H^{(k+1)}$ and α,β as indices of $\tilde{L}^{(k+1)}$. The universal rule for the relationship between the two sets of indices is

$$\binom{i}{j} = \binom{\alpha}{\beta} + k + 2.$$

We consider an arbitrary element, $\tilde{L}^{(k+1)}_{\alpha\beta}$. Due to the upper triangular shape of $\tilde{L}^{(k)}$, we have $\tilde{L}^{(k)}_{\alpha\beta}=0$ if $\alpha<\beta$. For $\alpha\geq\beta$, we plainly calculate, using the inductive hypothesis

$$\begin{aligned}\tilde{L}^{(k+1)}_{\alpha\beta} &= \sum_{q=1+\beta}^{1+\alpha} \tilde{L}^{(k)}_{1+\alpha,q}\cdot(\Delta^{(k+1)})^{-1}_{qq} \\ &= \sum_{q=1+\beta}^{1+\alpha}\prod_{\ell=1}^{k}(x_{k+2+\alpha}-x_{q+\ell})\cdot(x_{k+1+q}-x_q) \\ &= \prod_{\ell=1}^{k+1}(x_{k+2+\alpha}-x_{\beta+\ell})\cdot A = H^{(k)}_{ij}\cdot A,\end{aligned}$$

where A is the sum of terms that scales each summand to the desired quantity (by multiplying and dividing by missing factors). To complete the inductive proof, it suffices to show that $A=1$. It turns out that there are two main cases to consider, which we examine below.

Case 1. When $\alpha-\beta\leq k$, the term A can be expressed as

$$\begin{aligned}A =& \frac{x_{k+1+1+\beta}-x_{1+\beta}}{x_{k+2+\alpha}-x_{1+\beta}} + \frac{(x_{k+1+2+\beta}-x_{2+\beta})(x_{k+2+\alpha}-x_{k+1+1+\beta})}{(x_{k+2+\alpha}-x_{1+\beta})(x_{k+2+\alpha}-x_{2+\beta})} \\ &+\cdots+\frac{(x_{k+1+\gamma+\beta}-x_{\gamma+\beta})(x_{k+2+\alpha}-x_{k+2+\beta})\cdots(x_{k+2+\alpha}-x_{k+\gamma+\beta})}{(x_{k+2+\alpha}-x_{1+\beta})\cdots(x_{k+2+\alpha}-x_{\gamma-1+\beta})(x_{k+2+\alpha}-x_{\gamma+\beta})} \\ &+\cdots+\frac{(x_{k+1+\alpha}-x_{\alpha})(x_{k+2+\alpha}-x_{k+2+\beta})\cdots(x_{k+2+\alpha}-x_{k+1+\alpha})}{(x_{k+2+\alpha}-x_{1+\beta})\cdots(x_{k+2+\alpha}-x_{\alpha-1})(x_{k+2+\alpha}-x_{\alpha})} \\ &+\frac{\cancel{(x_{k+2+\alpha}-x_{1+\alpha})}(x_{k+2+\alpha}-x_{k+2+\beta})\cdots(x_{k+2+\alpha}-x_{k+1+\alpha})}{(x_{k+2+\alpha}-x_{1+\beta})\cdots(x_{k+2+\alpha}-x_{\alpha})\cancel{(x_{k+2+\alpha}-x_{1+\alpha})}}.\end{aligned}$$

Note that in the last term, the factor $(x_{k+2+\alpha}-x_{1+\alpha})$ in both the denominator and numerator cancels out, leaving the denominator to be the same as the second to last term. Combining the last two terms, we again get a common factor $(x_{k+2+\alpha}-x_\alpha)$ in denominator and numerator, which cancels out, and makes

the denominator of this term the same as that previous term. Continuing in this manner, we can recursively eliminate the terms from last to the first, leaving

$$\frac{\cancel{x_{k+2+\beta}} - x_{1+\beta} + x_{k+2+\alpha} - \cancel{x_{k+2+\beta}}}{x_{k+2+\alpha} - x_{1+\beta}} = 1.$$

In other words, we have shown that $A = 1$.

Case 2. When $\alpha - \beta \geq k + 1$, the denominators in terms of A will remain the same after they reach

$$(x_{k+2+\alpha} - x_{1+\beta}) \cdots (x_{k+2+\alpha} - x_{1+k+\beta}) = \prod_{\ell=1}^{k+1} (x_{k+2+\alpha} - x_{\beta+\ell}) := B.$$

Again, we begin by expressing A explicitly as

$$\begin{aligned}
A =& \frac{x_{k+1+1+\beta} - x_{1+\beta}}{x_{k+2+\alpha} - x_{1+\beta}} + \frac{(x_{k+1+2+\beta} - x_{2+\beta})(x_{k+2+\alpha} - x_{k+1+1+\beta})}{(x_{k+2+\alpha} - x_{1+\beta})(x_{k+2+\alpha} - x_{2+\beta})} \\
&+ \cdots + \frac{(x_{k+1+\gamma+\beta} - x_{\gamma+\beta})(x_{k+2+\alpha} - x_{k+2+\beta}) \cdots (x_{k+2+\alpha} - x_{k+\gamma+\beta})}{(x_{k+2+\alpha} - x_{1+\beta}) \cdots (x_{k+2+\alpha} - x_{\gamma-1+\beta})(x_{k+2+\alpha} - x_{\gamma+\beta})} \\
&+ \cdots + \frac{(x_{k+1+k+1+\beta} - x_{k+1+\beta})(x_{k+2+\alpha} - x_{k+2+\beta}) \cdots (x_{k+2+\alpha} - x_{k+k+1+\beta})}{(x_{k+2+\alpha} - x_{1+\beta}) \cdots (x_{k+2+\alpha} - x_{1+k+\beta})} \\
&+ \frac{(x_{k+1+k+2+\beta} - x_{k+2+\beta})(x_{k+2+\alpha} - x_{k+3+\beta}) \cdots (x_{k+2+\alpha} - x_{k+k+2+\beta})}{(x_{k+2+\alpha} - x_{1+\beta}) \cdots (x_{k+2+\alpha} - x_{1+k+\beta})} \\
&+ \cdots + \frac{(x_{k+1+\alpha} - x_{1+\alpha})(x_{k+2+\alpha} - x_{1+\alpha}) \cdots (x_{k+2+\alpha} - x_{k+\alpha})}{(x_{k+2+\alpha} - x_{1+\beta}) \cdots (x_{k+2+\alpha} - x_{1+k+\beta})} \\
&+ \frac{(x_{k+1+1+\alpha} - x_{1+\alpha})(x_{k+2+\alpha} - x_{2+\alpha}) \cdots (x_{k+2+\alpha} - x_{k+1+\alpha})}{(x_{k+2+\alpha} - x_{1+\beta}) \cdots (x_{k+2+\alpha} - x_{1+k+\beta})}.
\end{aligned}$$

Now we divide first factor of the transition term, in the third line above, into two halves by

$$x_{k+1+k+1+\beta} - x_{k+1+\beta} = (x_{k+2+\alpha} - x_{1+k+\beta}) + (x_{k+1+k+1+\beta} - x_{k+2+\alpha}).$$

The first half triggers the recursive reduction on the first k terms exactly as in the first case, so the sum of the first k terms equal to 1 and we get

$$\begin{aligned}
B(A-1) =& -(x_{k+2+\alpha} - x_{k+k+2+\beta})(x_{k+2+\alpha} - x_{k+2+\beta}) \cdots (x_{k+2+\alpha} - x_{k+k+1+\beta}) \\
&+ (x_{k+1+k+2+\beta} - x_{k+2+\beta})(x_{k+2+\alpha} - x_{k+3+\beta}) \cdots (x_{k+2+\alpha} - x_{k+k+2+\beta}) \\
&+ \cdots + (x_{k+1+\alpha} - x_{1+\alpha})(x_{k+2+\alpha} - x_{1+\alpha}) \cdots (x_{k+2+\alpha} - x_{k+\alpha}) \\
&+ (x_{k+1+1+\alpha} - x_{1+\alpha})(x_{k+2+\alpha} - x_{2+\alpha}) \cdots (x_{k+2+\alpha} - x_{k+1+\alpha}).
\end{aligned}$$

Now we can do a recursive reduction starting from the first two terms, the sum of which is

$$\begin{aligned}
&\Big[x_{k+1+k+2+\beta} - x_{k+2+\beta} - (x_{k+2+\alpha} - x_{k+2+\beta})\Big](x_{k+2+\alpha} - x_{k+3+\beta}) \cdots (x_{k+2+\alpha} - x_{k+k+2+\beta}) \\
=& -(x_{k+2+\alpha} - x_{k+1+k+2+\beta})(x_{k+2+\alpha} - x_{k+3+\beta}) \cdots (x_{k+2+\alpha} - x_{k+k+2+\beta})
\end{aligned}$$

This can be combined with the third term in a similar fashion and the recursion continues. At the end, we get

$$\begin{aligned}
B(A-1) =& -(x_{k+2+\alpha} - x_{k+1+\alpha})(x_{k+2+\alpha} - x_{1+\alpha}) \cdots (x_{k+2+\alpha} - x_{k+\alpha}) \\
&+ (x_{k+1+1+\alpha} - x_{1+\alpha})(x_{k+2+\alpha} - x_{2+\alpha}) \cdots (x_{k+2+\alpha} - x_{k+1+\alpha}) \\
=& \Big[x_{k+1+1+\alpha} - x_{1+\alpha} - (x_{k+2+\alpha} - x_{1+\alpha})\Big](x_{k+2+\alpha} - x_{2+\alpha}) \cdots (x_{k+2+\alpha} - x_{k+1+\alpha}) = 0.
\end{aligned}$$

That is, we have shown that $A = 1$.

With $A = 1$ proved between these two cases, we have completed the inductive argument, and hence the proof of the lemma.

A.2 Proof of Lemma 2 (inverse representation)

We prove Lemma 2, which claims that he inverse of falling factorial basis matrix is

$$(H^{(k)})^{-1} = \begin{bmatrix} C \\ \frac{1}{k!} \cdot D^{(k+1)} \end{bmatrix}, \tag{A.2}$$

where $D^{(k+1)}$ is the $(k+1)^{st}$ order discrete difference operator defined in (10), and the rows of the matrix $C \in \mathbb{R}^{(k+1) \times n}$ obey $C_1 = e_1$ and

$$C_{i+1} = \left[\frac{1}{i!} \cdot (\Delta^{(i)})^{-1} \cdot D^{(i)} \right]_1, \quad i = 1, \dots k.$$

Again we use induction on k. When $k = 0$, it is easily verified that

$$(H^{(0)})^{-1} = L_n^{-1} = \begin{bmatrix} e_1 \\ D^{(1)} \end{bmatrix} = \begin{bmatrix} e_1 \\ \frac{1}{0!} \cdot D^{(1)} \end{bmatrix}.$$

The rest of the inductive proof is relatively straightforward, following from Lemma 1, i.e., from (A.1). Inverting both sides of (A.1) gives

$$\begin{aligned} (H^{(k)})^{-1} &= \begin{bmatrix} I_k & 0 \\ 0 & L_{n-k} \end{bmatrix}^{-1} \cdot \begin{bmatrix} I_k & 0 \\ 0 & \Delta^{(k)} \end{bmatrix}^{-1} \cdot (H^{(k-1)})^{-1} \\ &= \begin{bmatrix} I_k & 0 \\ 0 & L_{n-k}^{-1} \end{bmatrix} \cdot \begin{bmatrix} I_k & 0 \\ 0 & (\Delta^{(k)})^{-1} \end{bmatrix} \cdot (H^{(k-1)})^{-1}. \end{aligned}$$

Now, using that $L_{n-k}^{-1} = \begin{bmatrix} e_1 \\ D^{(1)} \end{bmatrix}$, and assuming that $(H^{(k-1)})^{-1}$ obeys (A.2),

$$\begin{aligned} (H^{(k)})^{-1} &= \begin{bmatrix} I_k & 0 \\ 0 & \begin{bmatrix} e_1 \\ D^{(1)} \end{bmatrix} \end{bmatrix} \cdot \begin{bmatrix} I_k & 0 \\ 0 & (\Delta^{(k)})^{-1} \end{bmatrix} \cdot \begin{bmatrix} e_1 \\ \left[\frac{1}{1!}(\Delta^{(1)})^{-1} D^{(1)}\right]_1 \\ \vdots \\ \left[\frac{1}{(k-1)!}(\Delta^{(k-1)})^{-1} D^{(k-1)}\right]_1 \\ \frac{1}{(k-1)!} \cdot D^{(k)} \end{bmatrix} \\ &= \begin{bmatrix} e_1 \\ \left[\frac{1}{1!}(\Delta^{(1)})^{-1} D^{(1)}\right]_1 \\ \vdots \\ \left[\frac{1}{(k-1)!}(\Delta^{(k-1)})^{-1} D^{(k-1)}\right]_1 \\ \frac{1}{k!} \begin{bmatrix} e_1 \\ D^{(1)} \end{bmatrix} \cdot k(\Delta^{(k)})^{-1} \cdot D^{(k)} \end{bmatrix} = \begin{bmatrix} e_1 \\ \left[\frac{1}{1!}(\Delta^{(1)})^{-1} D^{(1)}\right]_1 \\ \vdots \\ \left[\frac{1}{(k-1)!}(\Delta^{(k-1)})^{-1} D^{(k-1)}\right]_1 \\ \left[\frac{1}{(k)!}(\Delta^{(k)})^{-1} D^{(k)}\right]_1 \\ \frac{1}{k!} \cdot D^{(k+1)} \end{bmatrix} = \begin{bmatrix} C \\ \frac{1}{k!} \cdot D^{(k+1)} \end{bmatrix}, \end{aligned}$$

as desired.

A.3 Algorithms for multiplication by $(H^{(k)})^T$ and $[(H^{(k)})^T]^{-1}$

Recall that, given a vector y, we write $y_{a:b}$ to denote its subvector $(y_a, y_{a+1}, \dots y_b)$, and we write cumsum and diff for the cumulative sum pairwise difference operators. Furthermore, we define flip to be the operator the reverses the order of its input, e.g., $\mathrm{flip}((1,2,3)) = (3,2,1)$, and we write $\circ$ to denote operator composition, e.g., $\mathrm{flip} \circ \mathrm{cumsum}$. The remaining two algorithms from Lemma 3 are given below, in Algorithms 3 and 4.

Algorithm 3 Multiplication by $(H^{(k)})^T$

Input: Vector to be multiplied $y \in \mathbb{R}^n$, order $k \geq 0$, sorted inputs vector $x \in \mathbb{R}^n$.
Output: y is overwritten by $(H^{(k)})^T y$.
for $i = 0$ to k **do**
  if $i \neq 0$ **then**
    $y_{(i+1):n} = y_{(i+1):n} ./ (x_{(i+1):n} - x_{1:(n-i)})$.
  end if
  $y_{(i+1):n} = \text{flip} \circ \text{cumsum} \circ \text{flip}(y_{(i+1):n})$.
end for
Return y.

Algorithm 4 Multiplication by $[(H^{(k)})^T]^{-1}$

Input: Vector to be multiplied $y \in \mathbb{R}^n$, order $k \geq 0$, sorted inputs vector $x \in \mathbb{R}^n$.
Output: y is overwritten by $[(H^{(k)})^T]^{-1} y$.
for $i = k$ to 0 **do**
  $y_{(i+1):n-1} = \text{flip} \circ \text{diff} \circ \text{flip}(y_{(i+1):n})$.
  if $i \neq 0$ **then**
    $y_{(i+1):n} = (x_{(i+1):n} - x_{1:(n-i)})^{-1} .* y_{(i+1):n}$.
  end if
end for
Return y.

A.4 Proof of Lemma 4 (proximity to truncated power basis)

Recall that we denote

$$\delta = \max_{i=1,\ldots n} (x_i - x_{i-1}),$$

and write $x_0 = 0$ for notational convenience. Taking the elementwise difference between the falling factorial and truncated power basis matrices, we get

$$H_{ij} - G_{ij} = \begin{cases} 0 & \text{for } i = 1, \ldots n,\ j = 1 \\ \prod_{\ell=1}^{j-1}(x_i - x_\ell) - x_i^{j-1} & \text{for } i > j-1,\ j = 2, \ldots k+1 \\ -x_i^{j-1} & \text{for } i \leq j-1,\ j = 2, \ldots k+1 \\ 0 & \text{for } i \leq j - \lceil k/2 \rceil,\ j \geq k+2 \\ -(x_i - x_{j-\lceil k/2 \rceil})^k & \text{for } j - \lceil k/2 \rceil < i \leq j-1,\ j \geq k+2 \\ \prod_{\ell=1}^{k}(x_i - x_{j-k-1+\ell}) - (x_i - x_{j-\lceil k/2 \rceil})^k & \text{for } i > j-1,\ j \geq k+2. \end{cases} \tag{A.3}$$

In the above, we use $\lceil z \rceil$ to denote the least integer greater than or equal to z (the ceiling function). We will bound the absolute value of each nonzero difference $H_{ij} - G_{ij}$ in (A.3). Starting with the second row,

$$\begin{aligned} \left| \prod_{\ell=1}^{j-1}(x_i - x_\ell) - x_i^{j-1} \right| &\leq x_i^{j-1} - (x_i - x_{j-1})^{j-1} \\ &= x_{j-1} \left[x_i^{j-2} + x_i^{j-3}(x_i - x_{j-1}) + \ldots + x_i(x_i - x_{j-1})^{j-3} + (x_i - x_{j-1})^{j-2} \right] \\ &\leq x_{j-1} \cdot (j-1) \cdot x_i^{j-2} \leq k\delta \cdot k \cdot 1 \leq k^2 \delta. \end{aligned}$$

In the second line above, we used the expansion

$$a^k - b^k = (a-b)(a^{k-1} + a^{k-2}b + \ldots + b^{k-1}), \tag{A.4}$$

and in the third line, we used the fact that $j-1 \leq k$, so that $x_{j-1} \leq k\delta$, and also $0 \leq x_i \leq 1$. The third row of (A.3) is simpler. Since $0 \leq x_i \leq 1$ and $i \leq j-1 < k$,

$$|-x_i^{j-1}| \leq x_i \leq k\delta.$$

For the fourth row in (A.3), using the range of i, j, and the fact that $k\delta \leq 1$,

$$|-(x_i - x_{j-\lceil k/2 \rceil})^k| \leq (x_{j-1} - x_{j-\lceil k/2 \rceil})^k \leq (k\delta)^k \leq k\delta.$$

This leaves us to deal with the last row in (A.3). Defining $p = i$, $q = j-(k+1)$, the problem transforms into bounding

$$\prod_{\ell=1}^{k} (x_p - x_{\ell+q}) - (x_p - x_{\lfloor \frac{k+2}{2} \rfloor + q})^k,$$

for any $p = k+2, k+3, \ldots n$, $q = 1, \ldots p-k$, where now $\lfloor z \rfloor$ denotes the greatest integer less than or equal to z (the floor function). We let $\mu_{pq} = x_p - x_{\lfloor \frac{k+2}{2} \rfloor + q}$ and $\eta_q = x_p - x_{q+1} - \mu_{pq}$. Note that η_q is the gap between the maximum multiplicant in the first term above and μ_{pq}. Then

$$\eta_q = x_{\lfloor \frac{k+2}{2} \rfloor + q} - x_{q+1} \leq k\delta.$$

Therefore

$$\begin{aligned}
\prod_{\ell=1}^{k} (x_p - x_{\ell+q}) - (x_p - x_{\lfloor \frac{k+2}{2} \rfloor + q})^k &\leq (x_p - x_{1+q})^k - \mu_{pq}^k \\
&= (\mu_{pq} + \eta_q)^k - \mu_{pq}^k \\
&= k\delta \cdot \sum_{\ell=0}^{k-1} (\mu_{pq} + \eta_q)^\ell \mu_{pq}^{k-\ell} \\
&\leq k^2\delta \cdot (\mu_{pq} + \eta_q)^k \leq k^2\delta.
\end{aligned}$$

The third line above follows again from the expansion (A.4), and the fact that $\eta_q \leq k\delta$. The fourth line uses $\mu_{pq} + \eta_q \geq \mu_{pq}$, and ultimately $\mu_{pq} + \eta_q = x_p - x_{1+q} \in [0, 1]$. This completes the proof.

A.5 Proof of Theorem 1 (trend filtering rate, fixed inputs)

This proof follows the same strategy as the convergence proofs in Tibshirani (2014). Recall that the trend filtering estimate (13) can be expressed in terms of the lasso problem (14), in that $\hat{\beta} = H^{(k)}\hat{\alpha}$; also consider consider the problem

$$\hat{\theta} = \operatorname*{argmin}_{\theta \in \mathbb{R}^n} \frac{1}{2}\|y - G^{(k)}\theta\|_2^2 + \lambda' \cdot \sum_{j=k+2}^{n} |\theta_j|, \tag{A.5}$$

where $G^{(k)}$ is the truncated power basis matrix of order k. Let $\mu = (f_0(x_1), \ldots f_0(x_n)) \in \mathbb{R}^n$ denote the true function evaluated across the inputs. Then under the assumptions of Theorem 1, it is known that

$$\|G^{(k)}\hat{\theta} - \mu\|_2^2 = O_{\mathbb{P}}(n^{-(2k+2)/(2k+3)}),$$

when $\lambda = \Theta(n^{1/(2k+3)})$; see Theorem 10 of Mammen & van de Geer (1997). It now suffices to show that $\|H^{(k)}\hat{\alpha} - G^{(k)}\hat{\theta}\|_2^2 = O_{\mathbb{P}}(n^{-(2k+2)/(2k+3)})$, since $\|H^{(k)}\hat{\alpha} - \mu\|_2^2 \leq 2\|H^{(k)}\hat{\alpha} - G^{(k)}\hat{\theta}\|_2^2 + 2\|G^{(k)}\hat{\theta} - \mu\|_2^2$. For this, we can use the results in Appendix B of Tibshirani (2014), specifically Corollary 4 of this work, to argue that we have $\|H^{(k)}\hat{\alpha} - G^{(k)}\hat{\theta}\|_2^2 = O_{\mathbb{P}}(n^{-(2k+2)/(2k+3)})$ as long as $\lambda = (1+\delta)\lambda'$ for any $\delta > 0$, and

$$n^{(2k+2)/(2k+3)} \cdot \max_{i,j=1,\ldots n} |G_{ij}^{(k)} - H_{ij}^{(k)}| \to 0 \quad \text{as } n \to \infty.$$

But by Lemma 4, and our condition (16) on the inputs, we have $\max_{i,j=1,\ldots n} |G_{ij}^{(k)} - H_{ij}^{(k)}| \leq k^2 \log n/n$, which verifies the above, and hence gives the result.

A.6 Proof of Lemma 5 (maximum gap between random inputs)

Given sorted i.i.d. draws $x_1 \leq \ldots \leq x_n$ from a continuous distribution supported on $[0,1]$, whose density is bounded below by $p_0 > 0$, we consider the maximum gap $\delta = \max_{i=1,\ldots n}(x_i - x_{i-1})$ (recall that we set $x_0 = 0$ for notational convenience). This is a well-studied quantity. In the case of a uniform distribution on $[0,1]$, we know that the spacings vector follows a symmetric Dirichelet distribution, which is equivalent to uniform sampling from an n-simplex, e.g., see David & Nagaraja (1970). Furthermore, the asymptotics of the kth largest gap have also been extensively studied, e.g., in Barbe (1992). Here, we provide a simple finite sample bound on δ, without using distributional or geometric characterizations, but rather a direct argument based on binning.

Consider an arbitrary point x in $[0, 1-\alpha]$. Then the probability that at least one draw from our underlying distribution occurs in $[x, x+\alpha]$ is bounded below by $1-(1-p_0\alpha)^n$. Now divide $[0,1]$ into bins of length α (the last bin can be overlapping with the second to last bin). Note that the event in which there is at least one sample point in each bin implies that the maximum gap δ between adjacent points is less than or equal to 2α. By the union bound, this event occurs with probability at least $1 - \lceil \frac{1}{\alpha} \rceil (1-p_0\alpha)^n$.

Let $\alpha = r \log n/(p_0 n)$, and assume n is sufficiently large so that $r \log n/(p_0 n) < 1$. Then we have

$$\begin{aligned}\Big\lceil \frac{1}{\alpha} \Big\rceil (1-p_0\alpha)^n &\leq \Big(\frac{1}{\alpha} + 1\Big)(1-p_0\alpha)^n = \frac{p_0 n + r\log n}{r \log n}\Big(1 - \frac{r \log n}{n}\Big)^n \\ &\leq 2p_0 n \exp(-r \log n) = 2p_0 n^{1-r}.\end{aligned}$$

Plugging in $r = 11$, we get the desired result for $C = 22$, i.e., with probability at least $1 - 2p_0 n^{-10}$, the maximum gap satisfies $\delta \leq 22 \log n/(p_0 n)$.

A.7 Proof of Corollary 1 (trend filtering rate, random inputs)

The proof of this result is entirely analogous to the proof of Theorem 1; the only difference is that

$$\max_{i=1,\ldots n-1} (x_{i+1} - x_i) = O_{\mathbb{P}}(\log n/n),$$

(i.e., convergence in probability now), and so accordingly,

$$n^{(2k+2)/(2k+3)} \cdot \max_{i,j=1,\ldots n} |G^{(k)}_{ij} - H^{(k)}_{ij}| \quad \xrightarrow{p} \quad 0 \quad \text{as } n \to \infty,$$

employing Lemmas 4 and 5. The same arguments now apply; the stability result in Corollary 4 in Appendix B of Tibshirani (2014) must now be applied to random predictor matrices, but this is an extension that is straightforward to verify.

B The higher order KS test

B.1 Motivating arguments

As described in the text, the classical KS test is

$$\mathrm{KS}(X_{(m)}, Y_{(n)}) = \max_{z_j \in Z_{(m+n)}} \left| \frac{1}{m}\sum_{i=1}^m 1\{x_i \leq z_j\} - \frac{1}{n}\sum_{i=1}^n 1\{y_i \leq z_j\} \right|, \tag{B.1}$$

over samples $X_{(m)} = (x_1, \ldots x_m)$ and $Y_{(n)} = (y_1, \ldots y_n)$, written in combined form as $Z_{(m+n)} = X_{(m)} \cup Y_{(n)} = (z_1, \ldots z_{m+n})$. It is well-known that the above definition is equivalent to

$$\mathrm{KS}(X_{(m)}, Y_{(n)}) = \max_{f\,:\,\mathrm{TV}(f) \leq 1} \left| \hat{\mathbb{E}}_{X_{(m)}}[f(X)] - \hat{\mathbb{E}}_{Y_{(n)}}[f(Y)] \right|, \tag{B.2}$$

where we write $\hat{\mathbb{E}}_{X_{(m)}}$ for the empirical expectation under $X_{(m)}$, so $\hat{\mathbb{E}}_{X_{(m)}}[f(X)] = 1/m \sum_{i=1}^m f(x_i)$, and similarly for $\hat{\mathbb{E}}_{Y_{(n)}}$. The equivalence between these two definitions follows from the fact that the maximum in (B.2) always occurs at an indicator function, $f(x) = 1\{x \leq z_i\}$, for some $i = 1, \ldots m+n$.

We now will step through a sequence of motivating arguments that lead to the definition of the higher order KS test in (20). The basic idea is to alter the constraint set in (B.2), and consider functions of bounded variation in their kth derivative, for some fixed $k \geq 0$. This gives

$$\max_{f\,:\,\mathrm{TV}(f^{(k)})\leq 1} \left|\hat{\mathbb{E}}_{X_{(m)}}[f(X)] - \hat{\mathbb{E}}_{Y_{(n)}}[f(Y)]\right|. \tag{B.3}$$

Is it possible to compute such a quantity? By a variational result in Mammen & van de Geer (1997), the maximum in (B.3) is always achieved by a kth order spline function. In principle, if we knew some finite set T containing the knots of the maximizing spline, then we could restrict our attention to the space of splines with knots in T. However, when $k \geq 2$, such a set T is not generically easy to find, because the knots of the maximizing spline in (B.3) can lie outside of the set of data samples $Z_{(m+n)} = \{z_1, \ldots z_{m+1}\}$ (Mammen & van de Geer, 1997). Therefore, we further restrict the functions in consideration in (B.3) to be kth order splines with knots contained in $Z = Z_{(m+n)}$. Letting $\mathcal{S}_Z^{(k)}$ denote the space of such spline functions, we hence examine

$$\max_{f\in\mathcal{S}_Z^{(k)}\,:\,\mathrm{TV}(f^{(k)})\leq 1} \left|\hat{\mathbb{E}}_{X_{(m)}}[f(X)] - \hat{\mathbb{E}}_{Y_{(n)}}[f(Y)]\right|. \tag{B.4}$$

(We have scaled the constrained set by $k!$ for convenience, which clearly will not affect the distribution of the test statistic.) As $\mathcal{S}_Z^{(k)}$ is a finite-dimensional function space (in fact, $(m+n)$-dimensional), we can rewrite (B.4) in a parametric form, similar to (B.1). Let $g_1, \ldots g_{m+n}$ denote the kth order truncated power basis with knots over the set of joined data samples Z. Then any function $f \in \mathcal{S}_Z^{(k)}$ with $\mathrm{TV}(f^{(k)}) \leq k!$ can be expressed as $f = \sum_{j=1}^{m+n} \alpha_j g_j$, where the coefficients satisfy $\sum_{j=k+2}^{m+n} |\alpha_j| \leq 1$. In terms of the evaluations of the function f over $z_1, \ldots z_{m+n}$, we have

$$\big(f(z_1), \ldots f(z_{m+n})\big) = G^{(k)}\alpha,$$

where $G^{(k)}$ is the truncated power basis matrix, i.e., its columns give the evaluations of $g_1, \ldots g_{m+n}$ over the points $z_1, \ldots z_{m+n}$. Therefore (B.4) can be re-expressed as

$$\max_{\sum_{j=k+2}^{m+n} |\alpha_j|\leq 1} \left|\frac{1}{m}\mathbb{1}_{X_{(m)}}^T G^{(k)}\alpha - \frac{1}{n}\mathbb{1}_{Y_{(n)}}^T G^{(k)}\alpha\right|. \tag{B.5}$$

Here $\mathbb{1}_{X_{(m)}}$ is an indicator vector of length $m+n$, indicating the membership of each point in the joined sample $Z_{(m+n)}$ to the set $X_{(m)}$. The analogous definition is made for $\mathbb{1}_{Y_{(n)}}$.

Upon inspection, some care must be taken in evaluating the maximum in (B.5). Let us decompose the coefficient vector into blocks as $\alpha = (\alpha_1, \alpha_2)$, where α_1 denotes the first $k+1$ coefficients and α_2 the last $m+n-k-1$. Then the constraint in (B.5) is simply $\|\alpha_2\|_1 \leq 1$, and it is not hard to see that since α_1 is unconstrained, we can choose it to make the criterion in (B.5) arbitrarily large. Therefore, in order to make (B.5) well-defined (finite), we employ the further restriction $\alpha_1 = 0$, yielding

$$\max_{\|\alpha_2\|_1\leq 1} \left|\frac{1}{m}\mathbb{1}_{X_{(m)}}^T G_2^{(k)}\alpha_2 - \frac{1}{n}\mathbb{1}_{Y_{(n)}}^T G_2^{(k)}\alpha_2\right|, \tag{B.6}$$

where $G_2^{(k)}$ denotes the last $m-n-k-1$ columns of $G^{(k)}$. A simple duality argument shows that (B.6) can be written in terms of the ℓ_∞ norm, finally giving

$$\mathrm{KS}_G^{(k)}(X_{(m)}, Y_{(n)}) = \left\|(G_2^{(k)})^T\left(\frac{\mathbb{1}_{X_{(m)}}}{m} - \frac{\mathbb{1}_{Y_{(n)}}}{n}\right)\right\|_\infty, \tag{B.7}$$

matching the our definition of the kth order KS test in (20). Note that when $k = 0$, this reduces to the usual (classic) KS test in (B.1).

For $k \geq 1$, unlike the usual KS test which requires $O(m+n)$ operations, the kth order KS test in (B.7) requires $O((m+n)^2)$ operations, due to the lower triangular nature of $G^{(k)}$. Armed with our falling factorial basis, we can approximate $\mathrm{KS}_G^{(k)}(X^m, Y^n)$ by

$$\mathrm{KS}_H^{(k)}(X_{(m)}, Y_{(n)}) = \left\| (H_2^{(k)})^T \left(\frac{\mathbb{1}_{X_{(m)}}}{m} - \frac{\mathbb{1}_{Y_{(n)}}}{n} \right) \right\|_\infty, \tag{B.8}$$

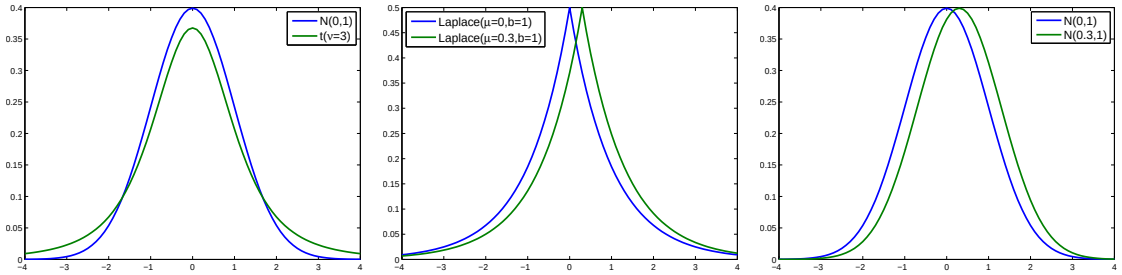
where $H^{(k)}$ is the kth order falling factorial basis matrix (and $H_2^{(k)}$ its last $m+n-k-1$ columns) over the points $z_1, \ldots z_{m+n}$. After sorting $z_1, \ldots z_{m+n}$, the statistic in (B.8) can be computed in $O((k+1)(m+n))$ time; see Algorithm 3, described above in Section A.3.

B.2 Additional experiments

In the main text, we presented two numerical experiments, on testing between samples from different distributions P, Q. In the first experiment $P = N(0,1)$ and $Q = t_3$, so the difference between P, Q was mainly in the tails; in the second, $P = \mathrm{Laplace}(0)$ and $Q = \mathrm{Laplace}(0.3)$, and the difference between P, Q was mainly in the centers of the distributions. The first experiment demonstrated that the power of the higher order KS test generally increased as we increased the polynomial degree k, the second demonstrated the opposite, i.e., that its power generally decreased for increasing k. Refer back to Figures 3 and 4 in the main text.

We should note that the first experiment was not carefully crafted in any way; the same performance is seen with a number of similar setups. However, we did have to look carefully to reveal the negative behavior shown in the second experiment. For example, in detecting the difference between mean-shifted standard normals (as opposed to Laplace distributions), the higher order KS tests do not encounter nearly as much difficulty. To demonstrate this, we examine a third experiment here with $P = N(0,1)$ and $Q = N(0.3,1)$. Figure B.1 gives a visual illustration of the distributions across the three experimental setups (the first two considered in the main text, and the third investigated here).

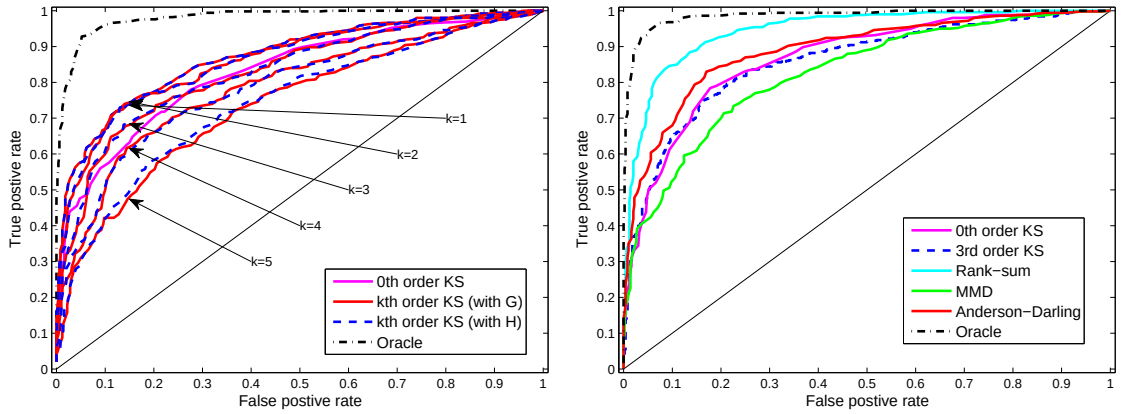
The ROC curves for experiment 3 are given in Figure B.2. The left panel shows that the test for $k = 1$ improves on the usual test ($k = 0$), even though the difference between the two distributions is mainly near their centers. The right panel shows that the higher order KS tests are competitive with other commonly used nonparametric tests in this setting. The results of this experiment hence suggest that the higher order KS tests provide a utility beyond simply detecting finer tail differences, and the tradeoff induced by varying the polynomial order k is not completely explained as a tradeoff between tail and center sensitivity.



(a) Experiment 1 (b) Experiment 2 (c) Experiment 3

Figure B.1: An illustration of distribution P vs. Q in our numerical experiments.

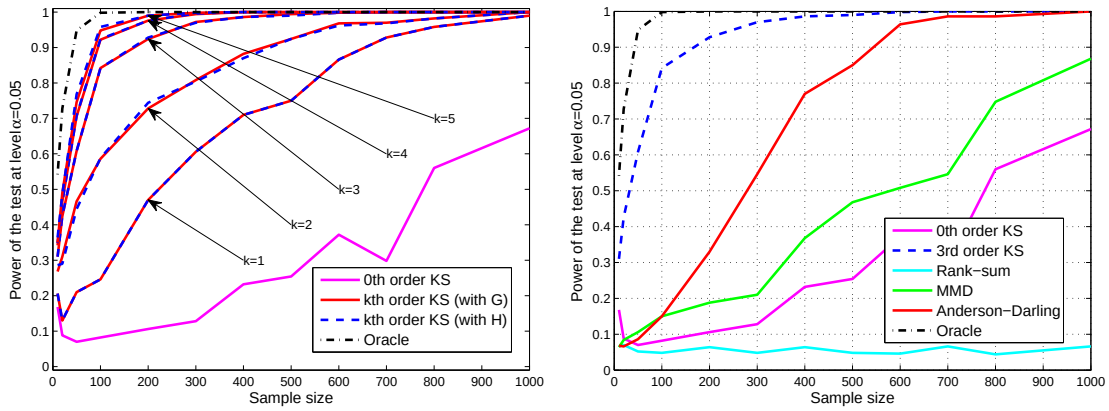
We also study the sample complexity of tests in the three experimental setups. Specifically, over $R = 1000$ repetitions, we find the true positive rate associated with a 0.05 false positive rate, as we let n vary over $10, 20, 50, 100, 200, \ldots 1000$. The results for this sample complexity sudy are shown in Figures B.3, B.4, and B.5. We see that the higher order KS tests perform quite favorably the first experimental setup, not so favorably in the second, and somewhere in the middle in the third.



(a) Comparing higher order KS tests

(b) Comparing other tests

Figure B.2: ROC curves for experiment 3, normal vs. shifted normal.



(a) Comparing higher order KS tests

(b) Comparing other tests

Figure B.3: Sample complexities at the level $\alpha = 0.05$ for experiment 1, normal vs. t.

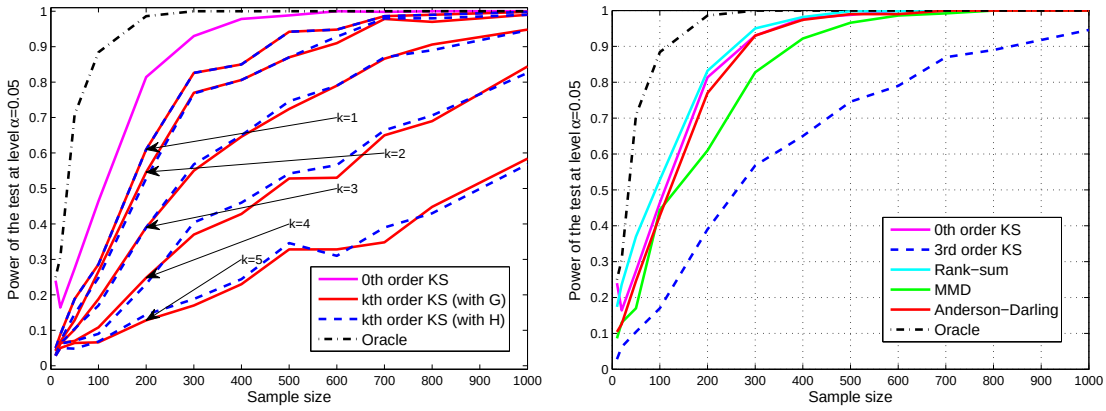
References

Barbe, Philippe. Limiting distribution of the maximal spacing when the density function admits a positive minimum. *Statistics & Probability Letters*, 14(1):53–60, 1992.

David, Herbert Aron and Nagaraja, Haikady Navada. *Order Statistics*. Wiley, Hoboken, 1970.

Mammen, Enno and van de Geer, Sara. Locally apadtive regression splines. *Annals of Statistics*, 25(1): 387–413, 1997.

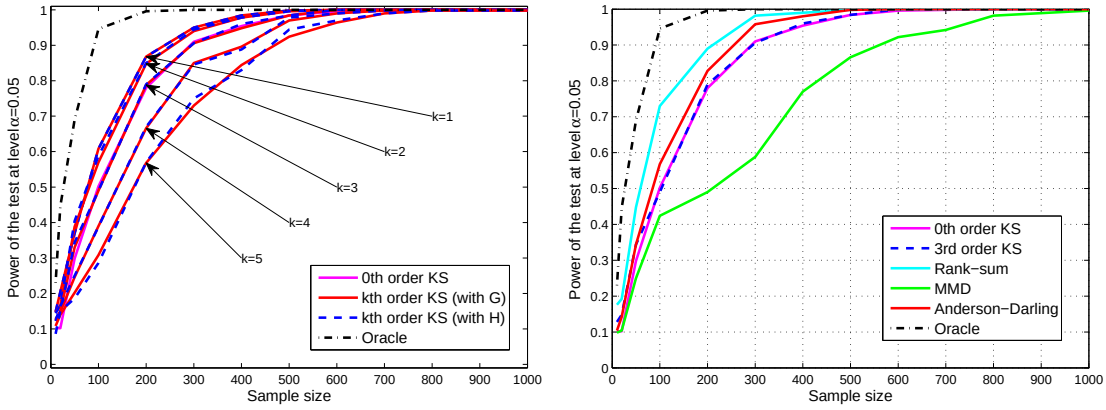
Tibshirani, Ryan J. Adaptive piecewise polynomial estimation via trend filtering. *Annals of Statistics*, 42(1): 285–323, 2014.



(a) Comparing higher order KS tests

(b) Comparing other tests

Figure B.4: Sample complexities at level $\alpha = 0.05$ in experiment 2, Laplace vs. shifted Laplace.



(a) Comparing higher order KS tests

(b) Comparing other tests

Figure B.5: Sample complexities at level $\alpha = 0.05$ in experiment 3, normal vs. shifted normal.