Appendix for Robust Hypothesis Test for Nonlinear Interaction in Nutrition-Environment Studies: A Kernel Ensemble Approach

Contents

Appendices

A Review: Kernel Machine Regression and Variance Component Test

We take the penalized likelihood approach to estimate parameters (β, h). Namely, we first specify $\mathscr{H}$ the candidate space and λ the penalty parameter, then estimate parameters $\hat{\theta} = (\hat{\beta}, \hat{h})$ by minimizing the penalized negative log likelihood:

$$(\hat{\beta}, \hat{h}) = \underset{\beta \in \mathbb{R},\, h \in \mathscr{H}}{argmin}\, L_\lambda(\beta, h), \qquad \text{where} \quad L_\lambda(\beta, h) = \sum_{i=1}^{n} ||y_i - \mathbf{x}_i\beta + h(z_i)||^2 + \lambda ||h||^2_{\mathscr{H}} \quad (1)$$

A.1 Estimating h using Kernel Ridge Regression

In principle, $\mathscr{H}$ needs to be rich enough to include the true function h, yet restrictive enough so that the optimization problem in (1) is computationally tractable. We strike this balance by characterizing h using kernel method (Schölkopf and Smola, 2002). Specifically, we assume $\mathscr{H}$ to be a Reproducing Kernel Hilbert Space (RKHS) generated by a user-defined, positive-definite kernel function $k(\mathbf{z}_i, \mathbf{z}_i')$, such that any $h \in \mathscr{H}$ can be expressed in terms the kernel function as $f(\mathbf{z}_i) = \langle f, k(\mathbf{z}_i, .) \rangle_{\mathscr{H}}$. The use of RKHS leads to a flexible yet computationally tractable way to model nonlinear h since, first, RKHS $\mathscr{H}$ can include rich class of functions through proper choice of its kernel function $k(\mathbf{z}, \mathbf{z}')$. For example, the space of linear functions can be generated using linear kernel $k(\mathbf{z}_i, \mathbf{z}_i') = \mathbf{z}_i^T \mathbf{z}_i'$, a space of twice-differentiable functions can be generated using Matérn 5/2 kernel $k(\mathbf{z}_i, \mathbf{z}_i') = \frac{2^{1-5/2}}{\Gamma(5/2)} \left(\frac{\sqrt{5||\mathbf{z}_i - \mathbf{z}_i||^2}}{\sigma^2} \right)^{\nu} K_\nu \left(\frac{\sqrt{5||\mathbf{z}_i - \mathbf{z}_i||^2}}{\sigma^2} \right)$, and a space of infinitely differentiable (therefore very smooth) functions can be generated using radial basis function (RBF) kernel $k(\mathbf{z}_i, \mathbf{z}_i') = exp(-\frac{1}{\sigma^2} * ||\mathbf{z}_i - \mathbf{z}_i'||^2_{\mathscr{H}})$. Second, by Representer theorem (Burges, 1999), solution to (1) can be expressed linearly as $h(\mathbf{z}_i) = \sum_{j=1}^{n} \alpha_i k(\mathbf{z}_i, \mathbf{z}_j)$. This property is the key to translate the difficult optimization problem in (1), which involves minimization over an potentially infinite-dimensional $\mathscr{H}$, to a tractable least-square problem that is finite-dimensional and has closed-form solution. Specifically, by plugging in the representor form into L_λ, we have:

$$L_\lambda(\beta, \{\alpha_i\}_{i=1}^n) = \sum_{i=1}^{n} [y_i - \mathbf{x}_i\beta + \sum_{j=1}^{n} \alpha_i * k(\mathbf{z}_i, \mathbf{z}_j)]^2 + \lambda \sum_{i=1}^{n} \sum_{j=1}^{n} [\alpha_i \alpha_j k(\mathbf{z}_i, \mathbf{z}_j)]$$

and if define $\mathbf{y}_{n\times 1} = [y_1, \ldots, y_n]^T$, $\mathbf{X}_{n\times p} = [\mathbf{x}_1^T, \ldots, \mathbf{x}_n^T]^T$, $\alpha = [\alpha_1, \ldots, \alpha_n]^T$ and also denote $\mathbf{K}_{n\times n}$ the kernel matrix with its $(i,j)^{th}$ element to be $\mathbf{K}_{i,j} = k(\mathbf{z}_i, \mathbf{z}_j)$, then (1) can be re-written as the

least-square problem:

$$(\hat{\beta}, \hat{\alpha}) = \underset{\beta \in \mathbb{R},\, \alpha \in \mathbb{R}^n}{argmin}\ L_\lambda(\beta, \alpha), \qquad \text{where } L_\lambda(\beta, \alpha) = ||\mathbf{y} - \mathbf{X}\beta - \mathbf{K}\alpha||^2 + \lambda\ \alpha^T \mathbf{K} \alpha \tag{2}$$

with solution:

$$\hat{\beta} = (\mathbf{X}^T(\lambda * \mathbf{I} + \mathbf{K})^{-1}\mathbf{X})^{-1}\mathbf{X}^T(\lambda * \mathbf{I} + \mathbf{K})^{-1}$$
$$\hat{\alpha} = (\mathbf{K} + \lambda \mathbf{I})^{-1}(\mathbf{y} - \mathbf{X}\hat{\beta})$$

This completes our description for the kernel-based estimation procedure for (β, α). Above procedure is commonly referred in the literature as *Kernel Machine Regression* (KMR) (Schölkopf and Smola, 2002).

Estimating Hyperparameters $k(\mathbf{z}, \mathbf{z}')$ and λ

So far we have assumed that both the kernel function $k(\mathbf{z}, \mathbf{z}')$ and the tuning parameter λ are known and fixed. In reality, however, (k, λ) are not known but have important consequences on the property of estimated $\hat{h}$. Specifically, the definition of k controls the **smoothness** of $\hat{h}$, which can be measured using integral of squared second derivatives $\int_{\mathbb{R}^q} [\nabla h(\mathbf{z})]^T \nabla h(\mathbf{z}) d\mathbf{z}$, and can intuitively understood as how "wiggly" the function is locally. On the other hand, the magnitude of λ controls the **norm** of $\hat{h}$, which can be measured using function norm $||h||_{\mathscr{H}}$, and can understood as how "varying" around zero the function is globally. In theory, regularization for $\hat{h}$ can proceed by controlling both its smoothness and its norm. However in practice, kernel method practitioners usually fix the choice of kernel function *a priori* (therefore fix the smoothness property of $\hat{h}$), and then control for the magnitude of $||\hat{h}||_{\mathscr{H}}$ by selecting λ adaptively from data.
Currently, there exists two popular approaches for λ selection: maximizing model-based likelihood (MLE), and minimizing cross-validation (CV) error. A classical example of the MLE approach within statistics literature is Liu et al. (2007), who argued that if denote $\tau = \frac{\sigma^2}{\lambda}$, then (2) can arise exactly from a linear mixed model (LMM):

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{h} + \varepsilon \qquad \text{where} \qquad \mathbf{h} \sim N(\mathbf{0}, \tau \mathbf{K}) \qquad \varepsilon \sim N(\mathbf{0}, \sigma^2 \mathbf{I}) \tag{3}$$

Therefore λ can be treated as part of the LMM's variance components parameters. If $\mathbf{K}$ is correctly specified, then unbiased estimates for variance components parameters can be obtained by maximizing the Restricted Maximum Likelihood (REML):

$$L_{RMLE}(\sigma, \tau) = -log|\mathbf{V}| - log|\mathbf{X}\mathbf{V}^{-1}\mathbf{X}^T| - (\mathbf{y} - \mathbf{X}\beta)^T \mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\beta)$$

where $\mathbf{V} = \tau\mathbf{K} + \sigma^2\mathbf{I}$. However, it is worth noting that REML is a model-based procedure. Therefore improper estimates for $\lambda = \frac{\sigma^2}{\tau}$ may arise when the family of kernel functions are mis-specified. On the contrary, λ estimates from CV-based approaches does not rely on assumptions of the model, and are more robust against model mis-specification (Wahba, 1990).

A.2 Testing for Nonlinear Effect using Variance Component Test

The KMR-LMM connection introduced in (3) opens up the arsenal of tools from Linear Mixed Model for inference tasks in Kernel Machine Regression. Most notably, (Maity and Lin, 2011) has utilized this connection to propose a recipe to construct a variance component test (Lin, 1997) for the following general null hypothesis:

$$H_0: \quad h \in \mathscr{H}_0 \tag{4}$$

where $\mathscr{H}_0$ is the space of functions under the null hypothesis. In this section, we will adapt this receipe for the current hypothesis of interest:

$$H_0: \quad h \in \mathscr{H}_{12}^{\perp}$$

A recipe for general hypothesis $h \in \mathscr{H}_0$

We first review the receipe presented in (Maity and Lin, 2011). Under the LMM formulation of Kernel Machine regression in (3), the authors assumed that h lies in a RKHS generated by a *garrote kernel function* $k_\delta(\mathbf{z},\mathbf{z}')$, which is constructed by attaching an extra *garrote parameter* δ to a regular kernel function. When $\delta = 0$, the garrote kernel function $k_0(\mathbf{z},\mathbf{z}') = k_\delta(\mathbf{z},\mathbf{z}')\Big|_{\delta=0}$ generates exactly $\mathscr{H}_0$ the space of functions under the null hypothesis. In order to adapt this recipe to his/her hypothesis of interest, practitioner only need to specify the form of the garrote kernel so that $\mathscr{H}_0$ corresponds to the null hypothesis. For example, if $k_\delta(\mathbf{z}) = k(\mathbf{z})^\delta$, $\delta = 0$ corresponds to the null hypothesis $H_0 : h(\mathbf{z}) = c$, i.e. $h(\mathbf{z})$ is a constant function. If $k_\delta(\mathbf{z}) = k(\delta * z_1, z_2, \ldots, z_q)$, $\delta = 0$ corresponds to the null hypothesis $H_0 : h(\mathbf{z}) = h(z_2, \ldots, z_q)$, i.e. the function $h(\mathbf{z})$ does not depend on z_1.
Given the definition of k_δ, this recipe constructs a testing procedure for $H_0 : h \in \mathscr{H}_0$ by considering the LMM formulation of the kernel machine regression under garrote kernel:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{h} + \varepsilon \qquad \text{where} \qquad \mathbf{h} \sim N(\mathbf{0}, \tau\mathbf{K}_\delta) \qquad \varepsilon \sim N(\mathbf{0}, \sigma^2\mathbf{I}) \tag{5}$$

where $\mathbf{K}_\delta$ is the kernel matrix generated by $k_\delta(\mathbf{z},\mathbf{z}')$. Therefore the garrote parameter δ can be treated as a variance component parameter in the linear mixed model. As a result, the general

hypothesis (4) is equivalent to:

$$H_0 : \delta = 0 \tag{6}$$

Maity and Lin (2011) proposed a REML-based score test for above hypothesis. We denote $\hat{\beta}$ and $\hat{\theta} = (\hat{\tau}, \hat{\sigma})$ the REML estimate for regression coefficients and the nuisance variance parameters under H_0. We also denote $\mathbf{K}_0$ the null kernel matrix whose $(i,j)^{th}$ entry is $k_\delta(\mathbf{z}, \mathbf{z}')\Big|_{\delta=0}$, and $\partial\mathbf{K}_0$ the null derivative kernel matrix whose $(i,j)^{th}$ entry is $\frac{\partial}{\partial\delta}k_\delta(\mathbf{z}, \mathbf{z}')\Big|_{\delta=0}$. The score test statistic for the hypothesis of interest is:

$$\hat{T}_0 = \hat{\tau} * (\mathbf{y} - \mathbf{X}\hat{\beta})^T \mathbf{V}_0^{-1} \partial\mathbf{K}_0 \mathbf{V}_0^{-1} (\mathbf{y} - \mathbf{X}\hat{\beta}) \tag{7}$$

where $\mathbf{V}_0 = \hat{\sigma}^2\mathbf{I} + \hat{\tau}\mathbf{K}_0$.
The null distribution of $\hat{T}$ is a mixture of chi-squares that can be approximated using a scaled chi-square distribution $\kappa\chi^2_\nu$ using Satterthwaite method (Zhang and Lin, 2003), i.e. we estimate (κ, ν) by matching the first two moments of T:

$$\begin{aligned} \kappa * \nu &= E(T) = \hat{\tau} * tr\Big(\mathbf{V}_0^{-1}\partial\mathbf{K}_0\Big) \\ 2 * \kappa^2 * \nu &= Var(T) = \hat{\mathbf{I}}_{\delta\delta} \end{aligned}$$

where $\hat{\mathbf{I}}_{\delta\delta} = \mathbf{I}_{\delta\delta} - \mathbf{I}_{\delta\theta}^T \mathbf{I}_{\theta\theta}^{-1} \mathbf{I}_{\delta\theta}$ is the efficient information of δ under REML. $\mathbf{I}_{\delta\delta}$, $\mathbf{I}_{\theta\theta}$ and $\mathbf{I}_{\delta\theta}$ are submatrices of the REML information matrix, whose $(i,j)^{th}$ element can be expressed as:

$$\mathbf{I}_{\theta_i\theta_j} = tr\Big(\mathbf{P}_0 \, (\frac{\partial}{\partial\theta_i}\mathbf{V}_0) \, \mathbf{P}_0 \, (\frac{\partial}{\partial\theta_j}\mathbf{V}_0)\Big)$$

where $\mathbf{P}_0 = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}^T(\mathbf{X}\mathbf{V}^{-1}\mathbf{X}^T)^{-1}\mathbf{X}\mathbf{V}^{-1}$ is the "scaled projection matrix" under REML, such that $\mathbf{P}_0\mathbf{y} = \frac{1}{\sigma^2}(\mathbf{y} - \mathbf{X}\hat{\beta} - \hat{\mathbf{h}})$ under the correct model. Consequently, the solution to above equations is:

$$\begin{aligned} \hat{\kappa} &= \hat{\mathbf{I}}_{\delta\theta} \Big/ \Big[\hat{\tau} * tr\Big(\mathbf{V}_0^{-1}\partial\mathbf{K}_0\Big)\Big] \\ \hat{\nu} &= \Big[\hat{\tau} * tr\Big(\mathbf{V}_0^{-1}\partial\mathbf{K}_0\Big)\Big]^2 / (2 * \hat{\mathbf{I}}_{\delta\theta}) \end{aligned}$$

Given test statistic $\hat{\mathbf{T}}$, we can compute the p-value of this test by examine the tail probability of $\hat{\kappa}\chi^2_{\hat{\nu}}$:

$$p = P(\hat{\kappa}\chi^2_{\hat{\nu}} > \hat{T}) = P(\chi^2_{\hat{\nu}} > \hat{T}/\hat{\kappa})$$

A.3 Testing for Kernel Interaction with Garrote Kernel

We now study how to adapt above recipe to the hypothesis of interest

$$H_0: \quad h \in \mathscr{H}_{12}^{\perp}$$

by designing a garrote kernel $k_\delta(\mathbf{z},\mathbf{z}')$ such that $k_0(\mathbf{z},\mathbf{z}')$ generates exactly the space of functions under H_0. We achieve this by first construct a reproducing kernel Hilbert space $\mathscr{H}_0$ that properly characterize $h(\mathbf{z}_i)$ under the null hypothesis, i.e. the space of functions not containing $(\mathbf{z}_{1,i},\mathbf{z}_{2,i})$ interaction, and then identify the kernel function that corresponds to the constructed RKHS.
We construct $\mathscr{H}_0$ by consider the tensor-product construction of RKHS on the product domain $(\mathbf{z}_{1,i},\mathbf{z}_{2,i}) \in \mathbb{R}^{q_1} \times \mathbb{R}^{q_2}$ (Gu, 2013) for its ability in explicitly characterizing the space of "pure interaction" functions. Let $\mathbf{1} = \{f | f \propto 1\}$ be the RKHS of constant functions with kernel function $k(\mathbf{z},\mathbf{z}') = 1$, and let $\mathscr{H}_m$ be the RKHS of centered functions (i.e. $\int f(\mathbf{z}_m)\, d\mathbf{z}_m = 0$) with domain on the m^{th} covariate set $\mathbf{z}_m$. Above individual spaces can be combined to constructed the tensor product space $\mathscr{H} = \otimes_{m=1}^{2}(\mathbf{1} \oplus \mathscr{H}_m)$. $\mathscr{H}$ describes the space of functions that depends jointly on $\{\mathbf{z}_1,\mathbf{z}_2,\mathbf{z}_3\}$, and adopts below orthogonal decomposition:

$$\begin{aligned}
\mathscr{H} &= \otimes_{m=1}^{2}(\mathbf{1} \oplus \mathscr{H}_m) = (\mathbf{1} \oplus \mathscr{H}_1) \otimes (\mathbf{1} \oplus \mathscr{H}_2) \\
&= \mathbf{1} \oplus \Big\{ \mathscr{H}_1 \oplus \mathscr{H}_2 \Big\} \oplus \Big\{ \mathscr{H}_1 \otimes \mathscr{H}_2 \Big\} \\
&= \mathbf{1} \oplus \mathscr{H}_{12}^{\perp} \oplus \mathscr{H}_{12}
\end{aligned}$$

where we have denoted the two non-constant components in above decomposition as $\mathscr{H}_{12}^{\perp}$ and $\mathscr{H}_{12}$, respectively. As shown, we have used the tensor-product construction to decompose $\mathscr{H} = \mathbf{1} \oplus \mathscr{H}_{12}^{\perp} \oplus \mathscr{H}_{12}$ into mutually orthogonal subspaces with desired interpretation: $\mathscr{H}_{12}^{\perp}$ is the space of functions that does not contain the $(\mathbf{z}_{1,i},\mathbf{z}_{2,i})$ interaction, and $\mathscr{H}_{12}^{\perp}$ is "pure interaction" space whose elements contain only the interaction effect between $(\mathbf{z}_{1,i},\mathbf{z}_{2,i})$, and not the main effect of $\{\mathbf{z}_1,\mathbf{z}_2\}$ or any other pairwise interactions. Furthermore, it can be shown that $\mathscr{H}_{12}^{\perp}$ and $\mathscr{H}_{12}$ are also reproducing kernel Hilbert spaces (Aronszajn, 1950), since they are constructed from basis RKHSs using direct sum and tensor product. To summarize, we have identified a reproducing kernel Hilbert space for $\mathscr{H}_0$ that has the desired interpretation:

$$\begin{aligned}
\mathscr{H}_0 &= \mathscr{H}_{12}^{\perp} \\
\mathscr{H}_a &= \mathscr{H}_{12}^{\perp} \oplus \mathscr{H}_{12}
\end{aligned}$$

It only left to identify a suitable garrote kernel $k_\delta(\mathbf{z},\mathbf{z}')$, such that $k_\delta(\mathbf{z},\mathbf{z}')$ is associated with $\mathscr{H}_0$ when $\delta = 0$, and associated with $\mathscr{H}_a$ when $\delta > 0$. To this end, we notice that both $\mathscr{H}_{12}^{\perp}$ and $\mathscr{H}_{12}$

are composite spaces built from basis RKHSs using direct sum and tensor product. If denote $k_m(\mathbf{z}_m, \mathbf{z}'_m)$ the reproducing kernel associated with $\mathscr{H}_m$, we can construct kernel functions for composite spaces $\mathscr{H}_0$ and $\mathscr{H}_a$ using below two facts (Aronszajn, 1950):

1. The tensor product $\mathscr{H}_1 \otimes \mathscr{H}_2$ possesses the reproducing kernel:

$$k([\mathbf{z}_1, \mathbf{z}_2], [\mathbf{z}'_1, \mathbf{z}'_2]) = k_1(\mathbf{z}_1, \mathbf{z}'_1) * k_2(\mathbf{z}_2, \mathbf{z}'_2)$$

2. The direct sum $\mathscr{H}_1 \oplus \mathscr{H}_2$, where $\mathscr{H}_1$ and $\mathscr{H}_2$ are mutually orthogonal, possesses the reproducing kernel:

$$k([\mathbf{z}_1, \mathbf{z}_2], [\mathbf{z}'_1, \mathbf{z}'_2]) = k_1(\mathbf{z}_1, \mathbf{z}'_1) + k_2(\mathbf{z}_2, \mathbf{z}'_2)$$

Combining the individual kernels $k_m(\mathbf{z}_m, \mathbf{z}'_m)$ according to above rule, the reproducing kernel for $\mathscr{H}_{12} = (\mathscr{H}_1 \otimes \mathscr{H}_2)$ is:

$$k_{12}(\mathbf{z}_i, \mathbf{z}'_i) = k_1(\mathbf{z}_1, \mathbf{z}'_1)\, k_2(\mathbf{z}_2, \mathbf{z}'_2) \tag{8}$$

similarly, the reproducing kernel for $\mathscr{H}_0 = \mathscr{H}_{12}^{\perp}$ is:

$$k_0(\mathbf{z}_i, \mathbf{z}'_i) = k_1(\mathbf{z}_1, \mathbf{z}'_1) + k_2(\mathbf{z}_2, \mathbf{z}'_2) \tag{9}$$

We are now ready to produce the form of garrote kernel using k_0 and k_{12}. Since $\mathscr{H}_a = \mathscr{H}_0 \oplus \mathscr{H}_{12}$, the garrote kernel $k_\delta(\mathbf{z}, \mathbf{z}')$ need to take below form under the null and the alternative hypothesis:

$$H_0 : k_\delta(\mathbf{z}, \mathbf{z}')\Big|_{\delta=0} = k_0(\mathbf{z}, \mathbf{z}')$$
$$H_a : k_\delta(\mathbf{z}, \mathbf{z}')\Big|_{\delta>0} = k_0(\mathbf{z}, \mathbf{z}') + k_{12}(\mathbf{z}, \mathbf{z}')$$

Therefore an intuitive choice of $k_\delta(\mathbf{z}, \mathbf{z}')$ is:

$$k_\delta(\mathbf{z}, \mathbf{z}') = k_0(\mathbf{z}, \mathbf{z}') + I(\delta > 0) * k_{12}(\mathbf{z}, \mathbf{z}')$$

Alternatively, since kernel functions are scale invariant (i.e. for $c \in (0, \infty)$, $c * k(\mathbf{z}, \mathbf{z}')$ represent the same RKHS as $k(\mathbf{z}, \mathbf{z}')$), it is equally valid to write the garrote kernel as:

$$k_\delta(\mathbf{z}, \mathbf{z}') = k_0(\mathbf{z}, \mathbf{z}') + \delta * k_{12}(\mathbf{z}, \mathbf{z}') \tag{10}$$

In this work, we choose to work with the form (10) so that δ can be treated as a continuous variance component parameter, hence drawing connection with the previous literature on classical variance component tests. However, the first garrote kernel is interesting in its own right, since it may induce a simple vs.simple hypothesis (i.e. whether $I(\delta > 0) = 1$) for which likelihood ratio test is the most powerful (although more computationally expensive). We left this direction for the future work.

Finally, using the chosen form of the garrote kernel function, the $(i, j)^{th}$ element of the null derivative kernel matrix $\mathbf{K}_0$ is $\frac{\partial}{\partial \delta} k_\delta(\mathbf{z}, \mathbf{z}') = k_{12}(\mathbf{z}, \mathbf{z}')$, i.e. the null derivative kernel matrix $\partial \mathbf{K}_0$ is simply the kernel matrix $\mathbf{K}_{12}$ that corresponds to the interaction space. Therefore our test statistic of interest in (7) is simplified to:

$$\hat{T}_0 = \hat{\tau} * (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T \mathbf{V}_0^{-1} \, \mathbf{K}_{12} \, \mathbf{V}_0^{-1} (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \tag{11}$$

A complete summary of the proposed testing procedure is available in Algorithm 1.

B Cross-validated Ensemble of Kernels

B.1 Derivation for Ensemble Kernel Matrix

Given the ensemble hat matrix $\widehat{\mathbf{A}}$ in Section 4, we consider how to identify the ensemble kernel matrix $\widehat{\mathbf{K}}$ by solving:

$$\widehat{\mathbf{K}}(\widehat{\mathbf{K}}+\lambda_{\mathbf{K}}\mathbf{I})^{-1}=\widehat{\mathbf{A}}.$$

Specifically, if denote $(\mathbf{U}_A,\mathbf{U}_K)$ and $(\{\delta_{A,k}\}_{k=1}^n,\{\delta_{K,k}\}_{k=1}^n)$ the eigenvector and eigenvalues of $\widehat{\mathbf{A}}$ and $\widehat{\mathbf{K}}$, respectively, then the above system reduces to:

$$\mathbf{U}_A diag\Big(\delta_{A,k}\Big)\mathbf{U}_A^T=\mathbf{U}_K diag\Big(\frac{\delta_{K,k}}{\delta_{K,k}+\lambda_{\mathbf{K}}}\Big)\mathbf{U}_K^T$$

and adopts closed form solution $\mathbf{U}_K=\mathbf{U}_A$ and $\delta_{K,k}=\lambda_{\mathbf{K}}\frac{\delta_{A,k}}{1-\delta_{A,k}}$. Therefore the ensemble kernel matrix $\widehat{\mathbf{K}}$ is estimated as:

$$\widehat{\mathbf{K}}=\lambda_{\mathbf{K}}*\mathbf{U}_A diag\Big(\frac{\delta_{A,k}}{1-\delta_{A,k}}\Big)\mathbf{U}_A^T.$$

Choice of ensemble tunning parameter $\lambda_{\mathbf{K}}$

Notice that we have left the "ensemble tunning parameter" $\lambda_{\mathbf{K}}$ unspecified. In practice, $\lambda_{\mathbf{K}}$ serves only as a constant scaling factor for the kernel matrix $\mathbf{K}$, whose exact value does not impact either the prediction or the p-value calculation, since both procedures are scale invariant with respect to the kernel matrix. Therefore it can be set to a value of our choice. One common choice for $\lambda_{\mathbf{K}}$ is to set $\lambda_{\mathbf{K}}=min\Big(1,(\sum_{k=1}^n\frac{\delta_{A,k}}{1-\delta_{A,k}})^{-1}\Big)$ such that $tr(\widehat{\mathbf{K}})\leq 1$, this is because the Rademacher complexity of the overall ensemble can be upper-bounded as a function of $tr(\widehat{\mathbf{K}})$ (Lanckriet et al., 2004). Another interesting choice is $\lambda_{\mathbf{K}}=O\Big(min(\{\hat{\lambda}_d\}_{d=1}^D)\Big)$. Intuitively, this means the tunning parameter for ensemble kernel matrix $\widehat{\mathbf{K}}$ should grow in the same rate as the tunning parameter for the best-peforming base kernel, as the ensemble kernel matrix is expected to perform as well or better than the best-performing base kernel. Further, as we will show in Lemma A.1, such choice provides guarantee on the generalization performance of the ensemble by bounding the ensemble kernel matrix's decay rate of the tail sum of the eigenvalues.

B.2 Proofs for Generalization Bound

Notation:

1. $\lambda = \{\hat{\lambda}_d\}_{d=1}^D$ the ordered set of estimated tunning parameters such that $\hat{\lambda}_1 > \cdots > \hat{\lambda}_D$
2. $\hat{\mathbf{u}} = \{\hat{u}_d\}_{d=1}^D$ the set of estimated weights for simplex ensemble, such that $\hat{\mathbf{u}} > 0, \mathbf{1}^T\hat{\mathbf{u}} = 1$.
3. $\delta_{\mathbf{M}} = \{\delta_{\mathbf{M},k}\}_{k=1}^n$ the ordered set of eigenvalues for a positive semi-definite matrix $\mathbf{M}_{n\times n}$, such that $\delta_{\mathbf{M},1} > \cdots > \delta_{\mathbf{M},n} \geq 0$
4. $S_{\mathbf{M}}(\theta) = \sum_{k>\theta} \delta_{\mathbf{M},k}$ the tail sum of eigenvalues of $\mathbf{M}$

Furthermore, we require below assumptions for the estimation procedure. Notice that these assumptions do not imposing restrictive assumption on data, they are merely important details already built into CVEK procedure but may otherwise be overlooked.
Assumptions:

A1 The kernel library has fixed size D. For each base kernel function $k_d \in \{k_d\}_{d=1}^D$, the generated kernel matrix $\mathbf{K}_d$ is positive semi-definite and are standardized by their respective trace $\mathbf{K}_d = \mathbf{K}_d/tr(\mathbf{K}_d)$.

A2 The estimated ensemble weights $\hat{\mathbf{u}}$ obey the simplex constraint: $\hat{\mathbf{u}} \in \Delta = \{\mathbf{u}|\mathbf{u} \geq 0, \mathbf{1}^T\mathbf{u} = 1\}$

A3 The tunning parameter for the ensemble kernel matrix satisfy $\frac{\lambda_{\mathbf{K}}}{min\{\hat{\lambda}_d\}_{d=1}^D} = O(1)$

Lemma A.1. ***(Decay Rate of Tail Sum of Eigenvalues for Ensemble Kernel Matrix)***
Assume

$$S_{\mathbf{K}_d}(\theta) = \mathscr{O}\Big(r_d(\theta)\Big) \qquad d \in \{1,\ldots,D\}$$

i.e. for each positive semi-definite, trace-normalized base kernel matrix $\mathbf{K}_d = \mathbf{K}_d/tr(\mathbf{K}_d)$, assume its tail sum of eigenvalues $S_{\mathbf{K}_d}(\theta)$ has decay rate $O(r_d(\theta))$, where r_d is a monotonically decreasing function of θ.
Then the decay rate of $S_{\widehat{\mathbf{K}}}(\theta)$, the tail sum of eigenvalues for ensemble kernel matrix $\widehat{\mathbf{K}}$, is bounded by:

$$S_{\widehat{\mathbf{K}}}(\theta) \leq \frac{\lambda_{\mathbf{K}}}{\hat{\lambda}_D}\mathscr{O}\Big(\sum_{d=1}^D \hat{u}_d r_d(\theta)\Big) \tag{12}$$

*In particular, if (**A3**) holds (i.e. $\lambda_{\mathbf{K}} = O\Big(min(\{\hat{\lambda}_d\}_{d=1}^{D})\Big)$), the upper bound on decay rate becomes:*

$$S_{\widehat{\mathbf{K}}}(\theta) \leq \mathscr{O}\Big(\sum_{d=1}^{D} \hat{u}_d r_d(\theta)\Big) \tag{13}$$

Proof. Denote $\delta_{\mathbf{K}}$, $\delta_{\mathbf{A}}$, $\delta_{\mathbf{A}_d}$, $\delta_{\mathbf{K}_d}$ the ordered sets of eigenvalues for ensemble kernel matrix $\mathbf{K}$, ensemble hat matrix $\mathbf{A}$, d^{th} base hat matrix $\mathbf{A}_d$, and d^{th} base kernel matrix $\mathbf{K}_d$. Notice that for the k^{th} eigenvalue ($k \in \{1,\ldots,n\}$), below conditions hold:

C1: $\delta_{A_d,k} = \frac{\delta_{K_d,k}}{\hat{\lambda}_d + \delta_{K_d,k}}$, since $\widehat{\mathbf{A}}_d = \mathbf{K}_d(\mathbf{K}_d + \hat{\lambda}_d \mathbf{I})^{-1}$

C2: $\delta_{A,k} \leq \sum_{d=1}^{D} \hat{u}_d \delta_{A_d,k}$, since $\widehat{\mathbf{A}} = \sum_{d=1}^{D} \hat{u}_d \widehat{\mathbf{A}}_d$ and by Weyl's inequality (Weyl, 1912)

C3: $\delta_{\widehat{\mathbf{K}},k} = \lambda_{\mathbf{K}} \frac{\delta_{A,k}}{1-\delta_{A,k}}$, since $\widehat{\mathbf{K}} = \lambda_{\mathbf{K}} * \mathbf{U}_A diag\Big(\frac{\delta_{A,k}}{1-\delta_{A,k}}\Big)\mathbf{U}_A^T$.

To study the tail sum of $\delta_{\mathbf{K}}$, we will first express $\delta_{\mathbf{K},k}$ in terms of the eigenvalues of individual kernel matrices $\delta_{\mathbf{K}_d,k}$. Recall $\hat{\lambda}_D$ is the smallest element of $\lambda = \{\hat{\lambda}_d\}_{d=1}^{D}$, and $\frac{\delta_{K_d,k}}{\hat{\lambda}_d + \delta_{K_d,k}}$ are concave functions for $d \in \{1,\ldots,K\}$. Then by **C1** and **C2**:

$$\begin{aligned}
\delta_{A,k} &\leq \sum_{d=1}^{D} \hat{u}_d \delta_{A_d,k} = \sum_{d=1}^{D} \hat{u}_d \frac{\delta_{K_d,k}}{\hat{\lambda}_d + \delta_{K_d,k}} && \text{(upper bound by replacing } \hat{\lambda}_d \text{ with } \hat{\lambda}_D\text{)} \\
&\leq \sum_{d=1}^{D} \hat{u}_d \frac{\delta_{K_d,k}}{\hat{\lambda}_D + \delta_{K_d,k}} && \text{(upper bound by concavity)} \\
&\leq \frac{\sum_{d=1}^{D} \hat{u}_d \delta_{K_d,k}}{\hat{\lambda}_D + \sum_{d=1}^{D} \hat{u}_d \delta_{K_d,k}}
\end{aligned} \tag{14}$$

Now notice in **C3**, $\frac{\delta_{A,k}}{1-\delta_{A,k}}$ is a convex, monotonically increasing function since $\delta_{A,k} \in [0,1)$, then

plugging (14) into **C3**, we have:

$$
\begin{aligned}
\delta_{\widehat{\mathbf{K}},k} &= \lambda_{\mathbf{K}} * \frac{\delta_{A,k}}{1-\delta_{A,k}} && \text{(by monotonicity)} \\
&\leq \lambda_{\mathbf{K}} * \Big(\frac{\sum_{d=1}^{D} \hat{u}_d \delta_{\mathbf{K}_d,k}}{\hat{\lambda}_D + \sum_{d=1}^{D} \hat{u}_d \delta_{\mathbf{K}_d,k}}\Big) \Big/ \Big(1 - \frac{\sum_{d=1}^{D} \hat{u}_d \delta_{\mathbf{K}_d,k}}{\hat{\lambda}_D + \sum_{d=1}^{D} \hat{u}_d \delta_{\mathbf{K}_d,k}}\Big) \\
&= \lambda_{\mathbf{K}} * \Big(\frac{\sum_{d=1}^{D} \hat{u}_d \delta_{\mathbf{K}_d,k}}{\hat{\lambda}_D + \sum_{d=1}^{D} \hat{u}_d \delta_{\mathbf{K}_d,k}}\Big) \Big/ \Big(\frac{\hat{\lambda}_D}{\hat{\lambda}_D + \sum_{d=1}^{D} \hat{u}_d \delta_{\mathbf{K}_d,k}}\Big) \\
&= \frac{\lambda_{\mathbf{K}}}{\hat{\lambda}_D} \sum_{d=1}^{D} \hat{u}_d \delta_{\mathbf{K}_d,k}
\end{aligned}
$$

It immediately follows that for $S_{\widehat{\mathbf{K}}}$ the tail sum of $\delta_{\widehat{\mathbf{K}}}$, below inequality holds:

$$
\begin{aligned}
S_{\widehat{\mathbf{K}}} &= \sum_{k>\theta} \delta_{\widehat{\mathbf{K}},k} \leq \sum_{k>\theta} \Big(\frac{\lambda_{\mathbf{K}}}{\hat{\lambda}_D} \sum_{d=1}^{D} \hat{u}_d \delta_{\mathbf{K}_d,k}\Big) \\
&\quad \text{(extract constant multiplier and switch order of summation):} \\
&= \frac{\lambda_{\mathbf{K}}}{\hat{\lambda}_D} \Big(\sum_{d=1}^{D} \hat{u}_d \big(\sum_{k>\theta} \delta_{\mathbf{K}_d,k}\big)\Big) \qquad \text{(recall } S_{\mathbf{K}_d}(\theta) = \textstyle\sum_{k>\theta} \delta_{\mathbf{K}_d,k}\text{):} \\
&= \frac{\lambda_{\mathbf{K}}}{\hat{\lambda}_D} * \sum_{d=1}^{D} \hat{u}_d S_{\mathbf{K}_d}(\theta)
\end{aligned}
$$

Now we are ready to study the asymptotic behavior of $S_{\widehat{\mathbf{K}}}$. Recall that $S_{\mathbf{K}_d}(\theta) = \mathscr{O}\Big(r_d(\theta)\Big)$, then by the property of big O notation:

$$
S_{\widehat{\mathbf{K}}} \leq \frac{\lambda_{\mathbf{K}}}{\hat{\lambda}_D} * \sum_{d=1}^{D} \hat{u}_d S_{\mathbf{K}_d}(\theta) = \frac{\lambda_{\mathbf{K}}}{\hat{\lambda}_D} * \mathscr{O}\Big(\sum_{d=1}^{D} \hat{u}_d r_d(\theta)\Big)
$$

which is exactly the expression in (12). Furthermore, if set $\frac{\lambda_{\mathbf{K}}}{\hat{\lambda}_D} = O(1)$ as in (**(A3)**), then:

$$
S_{\widehat{\mathbf{K}}} \leq \mathscr{O}\Big(\sum_{d=1}^{D} \hat{u}_d r_d(\theta)\Big)
$$

which is exactly the expression in (13). □

(Mendelson, 2003)
??
??
(Santin and Schaback, 2016)

C Additional Result: Interaction with Sub-mixtures

To further validate the observations made in joint-mixture analysis, we complete our study by conducting a targeted analysis by breaking down the As, Pb, Mn mixture into its two-pollutant and single-pollutant submixtures, and conduct hypothesis tests for the interaction between submixture and each nutrient groups, conditioning on the nuisance interaction for the other pollutants and nutrient groups. In this section, we report analysis result from CVKE-NN. Conclusions drawn from CVKE-RBF are similar and is reported in Appendix along with that of the iSKAT, GKM, and GE-spline.

As shown in Table 1, for two-pollutant mixtures, the analysis yielded suggestive evidence of interaction ($p <\approx 0.1$) for As-Pb and As-Mn mixtures with all nutrient groups except for the macronutrients. Similarly for the single pollutants, As and Mn yielded either suggestive or statistically significant evidence of interaction with all nutrient groups except for the macronutrients, confirming the observations made in the joint mixture study.

Toxin/Nutrient	macro	mineral	vitamin A	vitamin B	vitamin, other
As, Pb	0.1709	0.0904	0.0485	0.1102	0.0731
As, Mn	0.2505	0.0686	0.0849	0.1076	0.0533
Pb, Mn	0.7197	0.6389	0.1313	0.4875	0.5381
As	0.2371	0.0764	0.0758	0.1052	0.0474
Pb	0.6066	0.5091	0.3266	0.5561	0.6586
Mn	0.2190	0.3620	0.0660	0.3122	0.0242

Table 1: $p-value$ for Nutrient - Environment Interaction Test for all possible two-pollutant and single-pollutant sub-mixtures using CVKE-NN

References

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society 68*(3), 337–404.

Burges, C. J. C. (1999). Advances in Kernel Methods. pp. 89–116. Cambridge, MA, USA: MIT Press.

Gu, C. (2013, January). *Smoothing Spline ANOVA Models*. Springer Science & Business Media. Google-Books-ID: 5VxGAAAAQBAJ.

Lanckriet, G. R. G., N. Cristianini, P. Bartlett, L. E. Ghaoui, and M. I. Jordan (2004). Learning the Kernel Matrix with Semidefinite Programming. *Journal of Machine Learning Research 5*(Jan), 27–72.

Lin, X. (1997, June). Variance component testing in generalised linear models with random effects. *Biometrika 84*(2), 309–326.

Liu, D., X. Lin, and D. Ghosh (2007, December). Semiparametric Regression of Multidimensional Genetic Pathway Data: Least-Squares Kernel Machines and Linear Mixed Models. *Biometrics 63*(4), 1079–1088.

Maity, A. and X. Lin (2011, December). Powerful tests for detecting a gene effect in the presence of possible gene-gene interactions using garrote kernel machines. *Biometrics 67*(4), 1271–1284.

Mendelson, S. (2003). On the Performance of Kernel Classes. *Journal of Machine Learning Research 4*(Oct), 759–771.

Santin, G. and R. Schaback (2016, August). Approximation of eigenfunctions in kernel-based spaces. *Advances in Computational Mathematics 42*(4), 973–993.

Schölkopf, B. and A. J. Smola (2002, January). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press. Google-Books-ID: y8ORL3DWt4sC.

Wahba, G. (1990, September). *Spline Models for Observational Data*. SIAM. Google-Books-ID: ScRQJEETs0EC.

Weyl, H. (1912, December). Das asymptotische Verteilungsgesetz der Eigenwerte linearer partieller Differentialgleichungen (mit einer Anwendung auf die Theorie der Hohlraumstrahlung). *Mathematische Annalen 71*(4), 441–479.

Zhang, D. and X. Lin (2003, January). Hypothesis testing in semiparametric additive mixed models. *Biostatistics (Oxford, England) 4*(1), 57–74.