A Convergence of means and standard deviations

Uniform convergence of means, $|\hat{\eta}_k - \eta_k|$: We use a proof strategy related to that in [15], with two important differences: (1) the statistic is an average, and not a U-statistic, (2) the kernel family is $\mathcal{K}$ in (3). Given the boundedness assumptions on the kernels and coefficients defining $\mathcal{K}$, the largest change to the sum in (4) which could arise by replacing v_i by v_i' is $16DKm^{-1}$. By McDiarmid's Theorem [12], we have that with probability $1-\delta$,

$$\sup_{k\in\mathcal{K}} |\hat{\eta}_k - \eta_k| \leq \mathbf{E}_V \sup_{k\in\mathcal{K}} |\hat{\eta}_k - \eta_k| + 8DK\sqrt{2m^{-1}\log\delta^{-1}},$$

where $\mathbf{E}_V$ is the expectation over all of $\{v_i\}_{i=1}^{m/2}$. We next seek to bound the expectation on the right hand side. Using symmetrization,

$$\mathbf{E}_V \sup_{k\in\mathcal{K}} |\hat{\eta}_k - \eta_k| \leq 2\mathbf{E}_{V,\rho} \sup_{k\in\mathcal{K}} \left| \frac{2}{m} \sum_{i=1}^{m/2} \rho_i h_k(v_i) \right| =: R_{m/2}(\mathcal{K}, h_k),$$

where $\rho_i \in \{-1, 1\}$, each with probability $1/2$. The term $R_{m/2}(\mathcal{K}, h_k)$ is a Rademacher chaos complexity of order one. We now bound this quantity for the family $\mathcal{K}$ defined in (3). Note that a constraint on $\|\beta\|_1$ for this family is needed, since the normalization provided by σ_k has been omitted due to our bounding strategy. As a first step, rather than computing $R_{m/2}(\mathcal{K}, h_k)$, we use the larger class $\mathcal{K}'$ of kernels for which we omit the constraint $\beta \succeq 0$ and require $\|\beta\|_1 \leq D$, since by [3, Theorem 12(1)], $R_{m/2}(\mathcal{K}, h_k) \leq R_{m/2}(\mathcal{K}', h_k)$. This allows us to remove the absolute value sign in the Rademacher complexity. Next, define $g_i \in \mathcal{N}(0,1)$ to be independent standard Gaussian variables. By [3, Lemma 4], there exists an absolute constant C such that

$$\mathbf{E}_{V,\rho} \sup_{k\in\mathcal{K}'} \left(\frac{2}{m} \sum_{i=1}^{m/2} \rho_i h_k(v_i) \right) \leq C\mathbf{E}_{V,g} \sup_{k\in\mathcal{K}'} \left(\frac{2}{m} \sum_{i=1}^{m/2} g_i h_k(v_i) \right) =: CG_{m/2}(\mathcal{K}, h_k),$$

where $G_{m/2}(\mathcal{K})$ is the Gaussian complexity. We bound the latter using [3, Lemma 20]. Defining $z_u := \sum_{i=1}^{m/2} g_i h_u(v_i)$, then $\sup_{k\in\mathcal{K}'} \left(\sum_{i=1}^{m/2} g_i h_k(v_i) \right) = \max_{u\in\{1,\ldots,d\}} z_u$, and hence[2]

$$\begin{aligned} G_{m/2}(\mathcal{K}) = \mathbf{E}_g \max_{u\in\{1,\ldots,d\}} \left(\frac{2}{m} \sum_{i=1}^{m/2} g_i h_u(v_i) \right) &\leq \frac{2C}{m}\sqrt{\ln d} \max_{u,u'} \sqrt{\mathbf{E}_g (z_u - z_{u'})^2} \\ &= \frac{2C}{m}\sqrt{\ln d} \max_{u,u'} \sqrt{\mathbf{E}_g \left[\sum_{i=1}^{m/2} g_i \left(h_u(v_i) - h_{u'}(v_i) \right) \right]^2} \\ &= \frac{2C}{m}\sqrt{\ln d} \max_{u,u'} \sqrt{\sum_{i=1}^{m/2} \left(h_u(v_i) - h_{u'}(v_i) \right)^2} \leq \frac{C}{\sqrt{m}}\sqrt{\ln d}, \end{aligned}$$

where we use the boundedness of the k_u in the final line, and incorporate this upper bound into C. Combining the above inequalities yields that $\sup_{k\in\mathcal{K}} |\hat{\eta}_k - \eta_k| = O_P\left(m^{-1/2}\right)$.

Uniform convergence of standard deviations, $|\hat{\sigma}_k - \sigma_k|$: We begin with

$$\sup_{k\in\mathcal{K}} |\hat{\sigma}_{k,\lambda} - \sigma_{k,\lambda}| = \sup_{k\in\mathcal{K}} \frac{\left|\hat{\sigma}_k^2 - \sigma_k^2\right|}{|\hat{\sigma}_{k,\lambda} + \sigma_{k,\lambda}|} \leq \sup_{k\in\mathcal{K}} \sigma_k^{-1} \left|\hat{\sigma}_k^2 - \sigma_k^2\right| \leq C \sup_{k\in\mathcal{K}} \left|\hat{\sigma}_k^2 - \sigma_k^2\right|,$$

where we used the fact that σ_k is bounded away from zero for all $k \in \mathcal{K}$. Our goal now is to bound $\sup_{k\in\mathcal{K}} \left|\hat{\sigma}_k^2 - \sigma_k^2\right|$ in probability. We again make use of McDiarmid's inequality. The largest change

[2]The constant C below is not the same as that used earlier: we do not distinguish between such constants to simplify notation. In the same vein, C may change from line to line in the reasoning below.

to the sum in (7) which could arise by replacing w_i by w_i' is $(8)(8DK)^2 m^{-1} = 512D^2K^2m^{-1}$. Thus with probability $1-\delta$,

$$\sup_{k\in\mathcal{K}} \left|\hat{\sigma}_k^2 - \sigma_k^2\right| \leq \mathbf{E}_V \sup_{k\in\mathcal{K}} \left|\hat{\sigma}_k^2 - \sigma_k^2\right| + 16DK\sqrt{m^{-1}\log\delta^{-1}}.$$

Using symmetrization,

$$\mathbf{E}_V \sup_{k\in\mathcal{K}} \left|\hat{\sigma}_k^2 - \sigma_k^2\right| \leq 2\mathbf{E}_{W,\rho} \sup_{k\in\mathcal{K}} \left|\frac{4}{m}\sum_{i=1}^{m/4} \rho_i h_{\Delta,k}^2(w_i)\right|,$$

where $\mathbf{E}_W$ is the expectation over all of $\{w_i\}_{i=1}^{m/2}$. Next we note that over the range $[-8DK, 8DK]$ of $h_{\Delta,k}$, the function $\phi(h_{\Delta,k}) = h_{\Delta,k}^2$ has Lipschitz constant $16DK$ (since $\left|h_{\Delta,k}^2(w_1) - h_{\Delta,k}^2(w_2)\right| \leq 16DK\left|h_{\Delta,k}(w_1) - h_{\Delta,k}(w_2)\right|$), and $h_{\Delta,k}^2(0) = 0$. Thus, from [3, Lemma 12(4)],

$$\mathbf{E}_{W,\rho} \sup_{k\in\mathcal{K}} \left|\frac{4}{m}\sum_{i=1}^{m/4} \rho_i h_{\Delta,k}^2(w_i)\right| \leq (2)(16DK)\mathbf{E}_{W,\rho}\left|\frac{4}{m}\sum_{i=1}^{m/4} \rho_i h_{\Delta,k}(w_i)\right|.$$

With $\mathcal{K}$ defined in (3), and proceeding via [3, Lemma 4, Lemma 20] as before, we get $\mathbf{E}_{W,\rho}\left|\frac{4}{m}\sum_{i=1}^{m/4} \rho_i h_{\Delta,k}(w_i)\right| \leq \frac{C}{\sqrt{m}}\sqrt{\ln d}$ for an absolute constant C, which yields that $\sup_{k\in\mathcal{K}}\left|\hat{\sigma}_k - \sigma_k\right| = O_P\left(m^{-1/2}\right)$.

B Supplementary experiments

We provide three sets of supplementary experiments. In Section B.1, we compare our kernel selection strategy to alternative approaches on three simple synthetic benchmark problems. In Section B.2, we obtain the Type I error for all three datasets in the main document (Section 5), and investigate the distribution over kernels chosen by the various criteria under the null hypothesis, when p and q are identical. In Section B.3, we present two additional experiments in distinguishing amplitude modulated audio signals.

B.1 Detecting simple differences in three synthetic benchmarks

In our first supplementary synthetic benchmark, we compared samples from two multivariate Gaussian distributions with unit covariance matrices, where the means differed in one dimension only. In the second, we again compared two multivariate Gaussians, but this time with identical means in all dimensions, and variance that differed in a single dimension. In both cases, we considered dimensionality over the range $2^1, \ldots, 2^5$. In our third experiment, we used the benchmark data of [15]: one distribution was a univariate Gaussian, and the second was a univariate Gaussian with a sinusoidal perturbation of increasing frequency (where higher frequencies correspond to differences in distribution that are more difficult to detect).

We chose the base kernels $\{k_u\}_{u=1}^d$ in (3) to be Gaussian kernels with bandwidth varying between 2^{-10} and 2^8, with a multiplicative step-size of $2^{0.2}$. Results are reported in Figure 3. In the case of p and q with differing means, all four strategies yield very similar performance, while in the case of differing variances, *max-ratio* and *opt* have a statistically significant advantage over both *max-mmd* and l_2. In the sinusoidal perturbation data, the l_2 strategy has by far the highest Type II error, while the remaining methods perform similarly. We remark that while l_2 achieves a higher value of the MMD statistic in comparison to *max-mmd* (as the statistic is maximized over a larger set of kernels), this results in a significant deterioration of the Type II error performance, as no constraint on the variance $\hat{\sigma}_{k,\lambda}$ is imposed.

B.2 Investigation of Type I error, and kernel choice when the null hypothesis holds

In Figure 4, we plot the Type I error for the three benchmarks considered in Section 5. In all cases, samples from the null distribution were obtained by independently drawing each of the training and

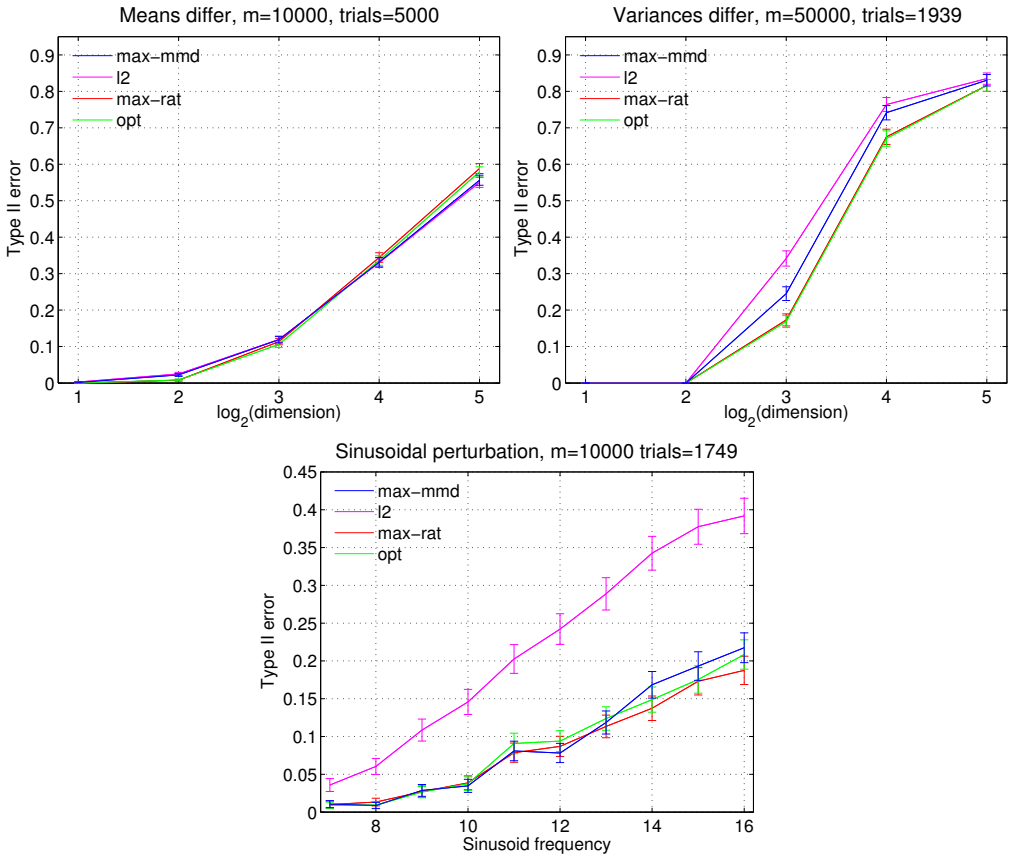

Figure 3: Type II error of various kernel selection strategies. **Left:** difference in means. **Right:** difference in variances. **Below:** sinusoidal difference. The test level was set to $\alpha = 0.05$. The error bars depict the 95% Wald confidence interval.

test points from p or q with equal probability. The Type I error was consistently close to or slightly below the design parameter of $\alpha = 0.05$ for all methods.

In Figure 5, we plot histograms of the kernels chosen for the three benchmarks in Section 5, under the null hypothesis. For methods l_2 and *opt* where β was non-zero over more than one kernel, fractional weights were assigned to the corresponding histogram bins. In the first experiment, we observe that the kernels are not biased towards particular features when the null hypothesis holds. In the second and third experiments, we note that under the null hypothesis, the kernel values are not clustered at the extremes of their allowed range.

B.3 Additional music experiments

We describe two additional music experiments. In the first, two Rammstein songs were compared (*Sehnsucht* vs *Engel*, from the album *Sehnsucht*), with parameters identical to the audio experiments in the main document, besides the setting $A = 0.3$. In the second experiment, two passages of contemporary jazz were compared (Christian Scott, *The Eraser* vs *KKPD*, from the album *Yesterday You Said Tomorrow*). Parameters were again identical to the earlier audio experiments, besides the setting $A = 0.7$. Results for both experiments are given in Figure 6.

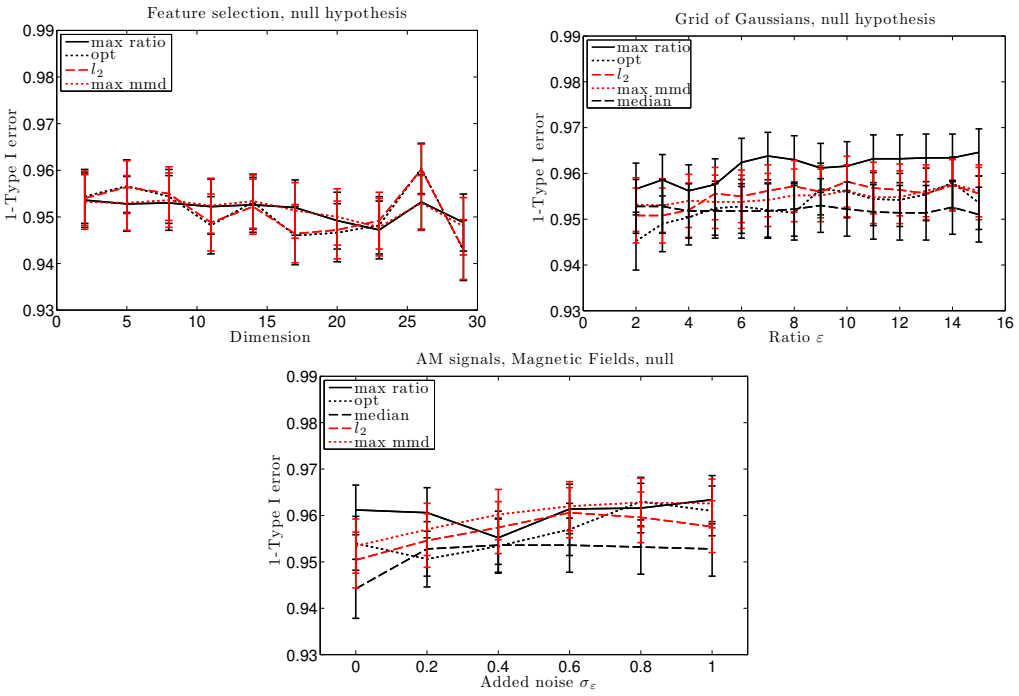

Figure 4: **Left:** Type I error for feature selection, **Right:** Type I error for grid of Gaussians. **Below:** Type I error for AM signals (Magnetic Fields sources). Average over 5000 trials, $m = n = 10^4$. The asymptotic test level was $\alpha = 0.05$. Error bars give the 95% Wald confidence interval.

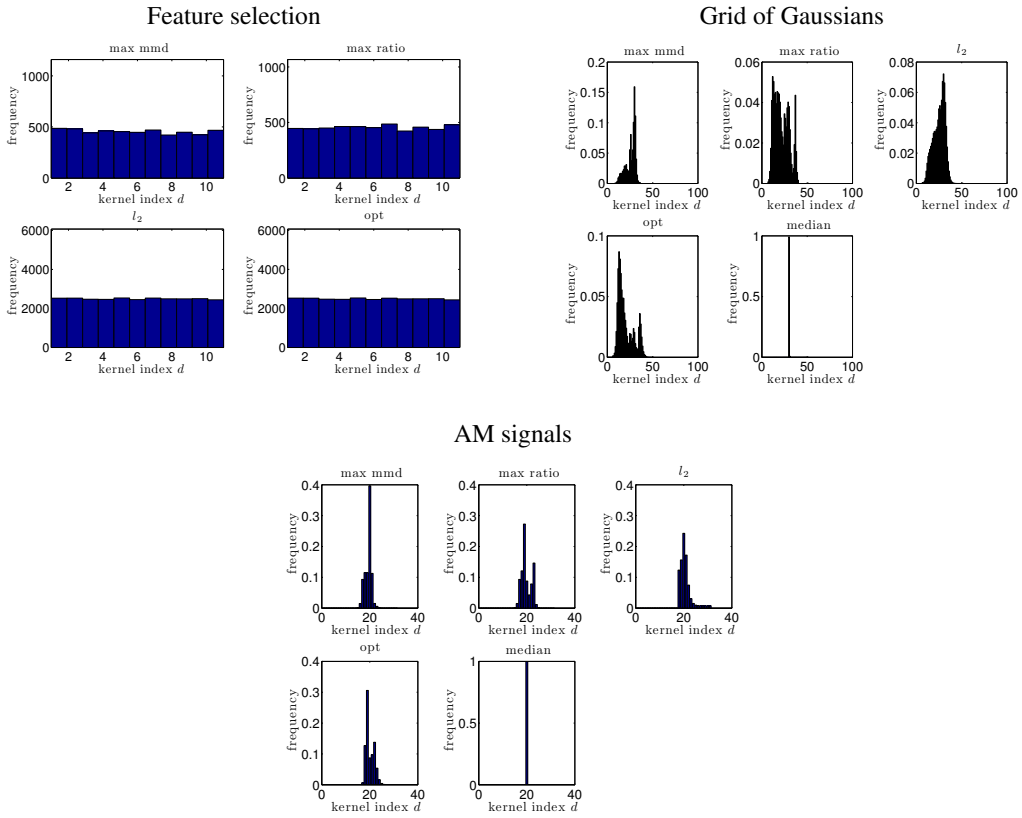

Figure 5: Kernels chosen when the null hypothesis holds, $p = q$. **Left:** feature selection in $d = 11$ dimensions, **Right:** grid of Gaussians, with ratio $\epsilon = 4$. **Below:** AM signals (Magnetic Fields sources), with added noise $\sigma_\varepsilon = 0.4$. Histograms were computed over 5000 trials, $m = 10^4$.

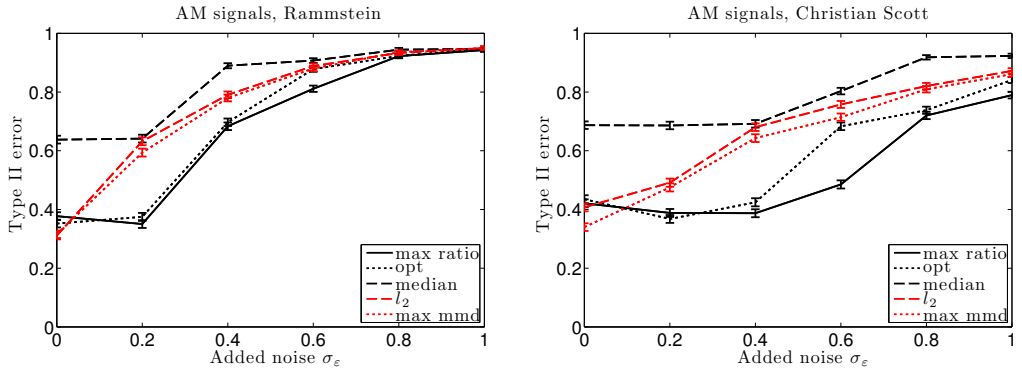

Figure 6: **Left:** AM results for Rammstein songs, **Right:** AM results for Christian Scott songs. Type II error vs added noise, average over 5000 trials, $m = n = 10^4$. The asymptotic test level was $\alpha = 0.05$. Error bars give the 95% Wald confidence interval.