Frequency Domain Predictive Modelling with Aggregated Data Supplementary Material

Avradeep Bhowmik
UT Austin
avradeep.1@utexas.edu

Joydeep Ghosh
UT Austin
ghosh@ece.utexas.edu

Oluwasanmi Koyejo
UIUC
sanmi@illinois.edu

I Frequency Domain Formulation

Consider our original loss function

$$\mathcal{L}(\boldsymbol{\beta}) = E[|\mathbf{x}(t)^\top \boldsymbol{\beta} - y(t)|^2]$$

As earlier, denote the residue term[1] at $\boldsymbol{\beta}$ as $\varepsilon_\beta(t) = \mathbf{x}(t)^\top \boldsymbol{\beta} - y(t)$, therefore our loss function can be written as

$$\mathcal{L}(\boldsymbol{\beta}) = E[\varepsilon_\beta(t)^2]$$

Suppose $P_{\varepsilon_\beta}(\omega)$ is the power spectral density of the residue term $\varepsilon_\beta(t)$. Then, we have

$$\mathcal{L}(\boldsymbol{\beta}) = E[\varepsilon_\beta(t)^2] = \int_{-\infty}^{\infty} P_{\varepsilon_\beta}(\omega) d\omega \tag{1}$$

As mentioned previously, we assume that $P_{\varepsilon_\beta}(\omega)$ decays rapidly with ω and almost vanishes beyond a certain $|\omega| > \omega_0$ (see section III for an extended discussion on this). Therefore, the integral on the right hand side can be approximated by a finite integral as

$$\int_{-\infty}^{\infty} P_{\varepsilon_\beta}(\omega) d\omega \approx \int_{-\omega_0}^{\omega_0} P_{\varepsilon_\beta}(\omega) d\omega$$

for a suitable ω_0.

Next, because we assume that $P_{\varepsilon_\beta}(\omega)$ exists finitely for every ω, the integral on the right hand side above can be approximated by averaging the readings of $P_{\varepsilon_\beta}(\omega)$ over a finite set of frequencies $\Omega = \{\omega_1, \omega_2, \cdots, \omega_M\}$ as

$$\int_{-\omega_0}^{\omega_0} P_{\varepsilon_\beta}(\omega) d\omega \approx \frac{1}{|\Omega|} \sum_{\omega \in \Omega} P_{\varepsilon_\beta}(\omega)$$

for a suitable Ω.

[1] Note that the residue process $\varepsilon_\beta(t)$ is equal to the error process $\epsilon(t)$ at $\boldsymbol{\beta} = \boldsymbol{\beta}^*$, where $\boldsymbol{\beta}^*$ is the true parameter

Finally, recall the definition of power spectral density

$$P_{\varepsilon_\beta}(\omega) = \lim_{T \uparrow \infty} E \left[\frac{\| \int_{-T}^{T} \varepsilon_\beta(t) e^{-\iota \omega t} dt \|^2}{2T} \right]$$

Again, because $P_{\varepsilon_\beta}(\omega)$ is assumed to exist finitely for every $\omega \in \Omega$, for a high enough T_0, the limit on the right hand side can be replaced by the value of the function at $T = T_0$

$$\lim_{T \uparrow \infty} E \left[\frac{\| \int_{-T}^{T} \varepsilon_\beta(t) e^{-\iota \omega t} dt \|^2}{2T} \right] \approx \frac{1}{2T_0} E\left[\|\mathcal{E}_{\beta, T_0}(\omega)\|^2 \right]$$

where $\mathcal{E}_{\beta,T_0}(\omega) = \mathbf{X}_{T_0}(\omega)\boldsymbol{\beta} - Y_{T_0}(\omega)$ is the T_0 restricted finite Fourier Transform of the residue at $\boldsymbol{\beta}$.

To summarize, the preceding discussion outlines the path by which our original loss function

$$\mathcal{L}(\boldsymbol{\beta}) = E[|\mathbf{x}(t)^\top \boldsymbol{\beta} - y(t)|^2]$$

can be substituted by an approximate frequency domain equivalent

$$\hat{\mathcal{L}}(\boldsymbol{\beta}) = \frac{1}{2T_0|\Omega|} \sum_{\omega \in \Omega} E\left[\|\mathbf{X}_{T_0}(\omega)\boldsymbol{\beta} - Y_{T_0}(\omega)\|^2 \right]$$

Since, the minimizer of an optimisation problem is invariant to positive scalar multiplication of the objective function, we use as our estimator

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{\omega \in \Omega} E\left[\|\mathbf{X}_{T_0}(\omega)\boldsymbol{\beta} - Y_{T_0}(\omega)\|^2 \right]$$

II Proofs of Main Results

We shall now make these ideas more concrete. We recall the main aspects of our setup below.

(i). We work with a parametric linear model, where the target variable $y(t)$ is regressod on predictor variables $\mathbf{x}(t)$ via a fixed parameter vector $\boldsymbol{\beta}^*$ as

$$y(t) = \mathbf{x}(t)^\top \boldsymbol{\beta}^* + \epsilon(t)$$

for each t.

The partial Fourier Transforms for our signals are

$$\mathbf{X}_T(\omega) = \int_{-T}^{T} \mathbf{x}(t) e^{-\iota\omega t} dt$$
$$Y_T(\omega) = \int_{-T}^{T} y(t) e^{-\iota\omega t} dt$$

(ii). We assume that each of our signals are weakly stationary stochastic processes with mean zero, and rapidly decaying autocorrelation function $\rho_{(\cdot)}(\tau)$ and finite variance $\rho_{(\cdot)}(0)$. In particular, this implies that $\varepsilon_\beta(t)$ is also centered and weakly stationary with rapidly decaying autocovariance function

We also assume finite power spectral density for all our signals, that is, we assume that

$$P_z(\omega) = \lim_{T\uparrow\infty} E\left[\frac{\|Z_T(\omega)\|^2}{2T}\right] \tag{2}$$
$$= \lim_{T\uparrow\infty} E\left[\frac{\|\int_{-T}^{T} z(t) e^{-\iota\omega t} dt\|^2}{2T}\right] \tag{3}$$
$$= \int_{-\infty}^{\infty} \rho_z(\tau) e^{-\iota\omega(\tau)} d\tau \tag{4}$$

is finite for every ω, and finitely integrable over $\omega \in (-\infty, \infty)$. It follows from these assumptions that the PSD will also be finite for the residue process $\varepsilon_\beta(t)$.

(iii). By linearity of Fourier transform, we have

$$Y_T(\omega) = \mathbf{X}_T(\omega)^\top \boldsymbol{\beta} + \mathcal{E}_T(\omega)$$

for any $T, \omega, \boldsymbol{\beta}$.

(iv). We define the optimal parameter $\boldsymbol{\beta}^*$ as the one that minimises the generalisation error, that is,

$$\boldsymbol{\beta}^* = \arg\min_{\boldsymbol{\beta}} \quad E\left[|\mathbf{x}(t)^\top \boldsymbol{\beta} - y(t)|^2\right] \tag{5}$$

We estimate our parameter in the frequency domain instead, as

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{\omega\in\Omega} E\left[\|\widehat{\mathbf{X}}_{T_0}(\omega)^\top \boldsymbol{\beta} - \widehat{Y}_{T_0}(\omega)\|^2\right]$$

for fixed parameters ω_0, T_0 and a set $\Omega = \{-\omega_0 < \omega_i < \omega_0 : i = 1, 2, \cdots |\Omega|\}$ of real valued "frequencies" sampled uniformly between $\omega \in (-\omega_0, \omega_0)$. Let $|\Omega| = D$. Also, define

$$\hat{\mathcal{L}}(\boldsymbol{\beta}) = \frac{1}{T_0|\Omega|} \sum_{\omega\in\Omega} E\left[\|\widehat{\mathbf{X}}_{T_0}(\omega)^\top \boldsymbol{\beta} - \widehat{Y}_{T_0}(\omega)\|^2\right]$$

We now prove some results that will be necessary in deriving our main theorems.

Lemma 1. *There exists an $0 < \xi_{\omega_0} < 1$ for every ω_0 (conversely, for every $\xi_{\omega_0} \in (0,1)$, there exists ω_0) such that*

$$(1 - \xi_{\omega_0}) E\left[|\mathbf{x}(t)^\top \boldsymbol{\beta} - y(t)|^2\right] \leq \int_{-\omega_0}^{\omega_0} P_{\varepsilon_\beta}(\omega) d\omega \leq E\left[|\mathbf{x}(t)^\top \boldsymbol{\beta} - y(t)|^2\right] \tag{6}$$

Proof. We use the following standard result. For any weakly stationary signal $\mathbf{z}$, we have

$$E\left[|z(t)|^2\right] = \int_{-\infty}^{\infty} P_z(\omega) d\omega \tag{7}$$

By equation (7), we have

$$E\left[|\mathbf{x}(t)^\top \boldsymbol{\beta} - y(t)|^2\right] = E[|\varepsilon_\beta(t)|^2] = \int_{-\infty}^{\infty} P_{\varepsilon_\beta}(\omega) d\omega$$

Also, by assumption, $P_{\varepsilon_\beta}(\omega)$ is finite for every ω, and finitely integrable over $(-\infty, \infty)$. Moreover, by definition of power spectral density, $P_{\varepsilon_\beta}(\omega) \geq 0$ for each ω. Hence, the result. ☐

The constant ξ_{ω_0} depends on the exact functional form of P_{ε_β}, or equivalently, of ρ. Standard rates can be obtained by using the fact that $\frac{P_{\varepsilon_\beta}(\omega)}{\int_{-\infty}^{\infty} P_{\varepsilon_\beta}(\omega) d\omega}$ is a valid probability density function, and using the tail probability results for the corresponding probability distribution.

For example, if ρ exhibits a Gaussian decay (analogous to normal distribution), that is, $\rho(\tau) \sim exp(-O(\tau^2))$, then P_{ε_β} also exhibits a Gaussian decay, that is $P_{\varepsilon_\beta}(\omega) \sim exp(-O(\omega^2))$, and therefore, $\xi_{\omega_0} \sim exp(-O(\omega_0^2))$. Similarly, if ρ exhibits power law/ Lorentzian decay (analogous to Cauchy distribution), that is, $\rho(\tau) \sim \frac{1}{O(\tau^2)}$, then P_{ε_β} exhibits exponential decay (Laplace distribution), that is $P_{\varepsilon_\beta}(\omega) \sim exp(-O(|\omega|))$, and therefore

$\xi_{\omega_0} \sim exp(-O(|\omega_0|)$. Similar arguments can be made for other decay rates using Fourier duality.

This makes intuitive sense because the more spread out $\rho(\tau)$ is, the more "peaky" $P_{\varepsilon_\beta}(\omega)$ is and the smaller the value of ω_0 required. This means that if the error terms are well-correlated, most of the instantaneous power will be concentrated within a very small range of frequencies.

Lemma 2. *Suppose* $\Omega = \{-\omega_0 < \omega_i < \omega_0 : i = 1, 2, \cdots |\Omega|\}$ *is a set of real valued "frequencies" sampled uniformly between* $[-\omega_0, \omega_0]$. *Then, for any* $\xi_D \in (0, 1)$, *with probability at least* $1 - exp(-O(|\Omega|^2\xi_m^2))$ *we have*

$$\int_{-\omega_0}^{\omega_0} P_{\varepsilon_\beta}(\omega)d\omega - \xi_D \le \frac{1}{|\Omega|}\sum_{\omega\in\Omega} P_{\varepsilon_\beta}(\omega) \le \int_{-\omega_0}^{\omega_0} P_{\varepsilon_\beta}(\omega)d\omega + \xi_D \tag{8}$$

Proof. This is standard Monte Carlo approximation. In particular, consider ω to be a random variable distributed uniformly in $(-\omega_0, \omega_0)$. Now consider the random variable $\zeta(\omega) = P_{\varepsilon_\beta}(\omega)$. Then, we have for this random variable, $\int_{-\omega_0}^{\omega_0} P_{\varepsilon_\beta}(\omega)d\omega = E_{U(-\omega_0,\omega_0)}[\zeta(\omega)] = E[\zeta]$.

Since P_{ε_β} is finite by our assumption, and ω has a finite support $(-\omega_0, \omega_0)$, we also have that $\zeta(\omega) = P_{\varepsilon_\beta}(\omega)$ has a finite support, and we have our result using Hoeffding's inequality[1]. □

Lemma 3. *Let* $\xi_{T_0} \in (0, 1)$. *Then, for every* ω, *there exists a* $T_0(\omega)$ *such that*

$$-\xi_{T_0} + P_{\varepsilon_\beta}(\omega) < \frac{1}{2T_0(\omega)} E\left[\|\mathbf{X}_{T_0}(\omega)^\top \boldsymbol{\beta} - Y_{T_0}(\omega)\|^2\right] < P_{\varepsilon_\beta}(\omega) + \xi_{T_0} \tag{9}$$

Proof. Define the partial power spectral density of $\varepsilon_\beta(t)$ as

$$g_{\varepsilon_\beta}(T;\omega) = E\left[\frac{\|\int_{-T}^{T} \varepsilon_\beta(t)e^{-\iota\omega t}dt\|^2}{2T}\right]$$

By definition of power spectral density, we have

$$P_{\varepsilon_\beta}(\omega) = \lim_{T\uparrow\infty} E\left[\frac{\|\int_{-T}^{T} \varepsilon_\beta(t)e^{-\iota\omega t}dt\|^2}{2T}\right] = \lim_{T\uparrow\infty} g_{\varepsilon_\beta}(T;\omega)$$

By assumption, the power spectral density is finite and converges for each ω to $P_{\varepsilon_\beta}(\omega)$. Therefore, for every ω, we have that $g_{\varepsilon_\beta}(T;\omega)$ must be a Cauchy sequence with respect to T. That is, for every $\omega \in (-\omega_0, \omega_0)$, and every $\xi_{T_0} \in (0, 1)$, $\exists T_0(\omega)$ such that for all $T > T_0(\omega)$,

$$-\xi_{T_0} + P_{\varepsilon_\beta}(\omega) < g_{\varepsilon_\beta}(T;\omega) < P_{\varepsilon_\beta}(\omega) + \xi_{T_0}$$

□

Remark: The exact value of T_0 does not contribute to computation time or space complexity, etc. beyond the computation of the respective Fourier Transforms, and can be chosen as large as required without any additional expenditure in the algorithm. In fact, the optimisation step itself does not depend on T_0, therefore by taking a large enough T_0, we can push ξ_{T_0} to as small as required.

IIa Proof of Theorem 1

We are now in a position to prove our first main result.

Proof: For any ξ_{T_0}, there always exists a T_0 that is the maximum $T_0(\omega)$ over all $\omega \in (-\omega_0, \omega_0)$ such that Lemma 3 is satisfied, i.e.,

$$\begin{aligned} T_0 = &\min\ T \\ s.t.\quad &|g_{\varepsilon_\beta}(T';\omega) - P_{\varepsilon_\beta}(\omega)| < \xi_{T_0} \\ &\forall T' > T, \forall\omega \in (-\omega_0, \omega_0) \end{aligned}$$

Combining Lemmata 1, 2 and 3 we have, for every $\xi_{T_0}, \xi_D, \xi_{\omega_0} \in (0, 1)$, there exist T_0, ω_0 such that for some set $\Omega = \{-\omega_0 < \omega_i < \omega_0 : i = 1, 2, \cdots |\Omega|\}$ sampled uniformly between $(-\omega_0, \omega_0)$, we have with probability at least $1 - exp(-O(|\Omega|^2\xi_m^2))$

$$\begin{aligned} &-\xi_{T_0} - \xi_D + (1 - \xi_{\omega_0})E\left[|\mathbf{x}(t)^\top \boldsymbol{\beta} - y(t)|^2\right] \\ &\le \frac{1}{2|\Omega|T_0}\sum_{\omega\in\Omega} E\left[\|\mathbf{X}_{T_0}(\omega)^\top \boldsymbol{\beta} - Y_{T_0}(\omega)\|^2\right] \\ &\le E\left[|\mathbf{x}(t)^\top \boldsymbol{\beta} - y(t)|^2\right] + \xi_D + \xi_{T_0} \end{aligned} \tag{10}$$

In other words,

$$-\xi_{T_0} - \xi_D + (1 - \xi_{\omega_0})\mathcal{L}(\boldsymbol{\beta}) \le \hat{\mathcal{L}}(\boldsymbol{\beta}; \omega_0, T_0, \Omega) \le \mathcal{L}(\boldsymbol{\beta}) + \xi_D + \xi_{T_0} \tag{11}$$

With some algebra, we have,

$$\begin{aligned} \mathcal{L}(\widehat{\boldsymbol{\beta}}) &< \left(\frac{1}{1-\xi_{\omega_0}}\right)\hat{\mathcal{L}}(\boldsymbol{\beta}; \omega_0, T_0, \Omega) + \frac{1}{1-\xi_{\omega_0}}(\xi_D + \xi_{T_0}) \\ &< \left(\frac{1}{1-\xi_{\omega_0}}\right)\hat{\mathcal{L}}(\boldsymbol{\beta}^*; \omega_0, T_0, \Omega) + \frac{1}{1-\xi_{\omega_0}}(\xi_D + \xi_{T_0}) \\ &< \left(\frac{1}{1-\xi_{\omega_0}}\right)(\mathcal{L}(\boldsymbol{\beta}^*) + \xi_D + \xi_{T_0}) + \frac{1}{1-\xi_{\omega_0}}(\xi_D + \xi_{T_0}) \\ &< \left(\frac{1}{1-\xi_{\omega_0}}\right)\mathcal{L}(\boldsymbol{\beta}^*) + \frac{2}{1-\xi_{\omega_0}}(\xi_D + \xi_{T_0}) \end{aligned}$$

where the first inequality is due to eq. (11), the second by definition of $\boldsymbol{\beta}^*$ and $\widehat{\boldsymbol{\beta}}$,and the final two by eq. (11). Therefore, we have

$$E\left[|\mathbf{x}(t)^\top\widehat{\boldsymbol{\beta}} - y(t)|^2\right] < \left(\frac{1}{1-\xi_{\omega_0}}\right) E\left[|\mathbf{x}(t)^\top\boldsymbol{\beta}^* - y(t)|^2\right] + \left(\frac{2}{1-\xi_{\omega_0}}\right)(\xi_D + \xi_{T_0}) \tag{12}$$

Choosing T_0, ω_0 and $|\Omega| = D$ such that $\xi_1 = \frac{\xi_{\omega_0}}{1-\xi_{\omega_0}}$, $\xi_2 = 2(\xi_D + \xi_{T_0})$ completes the proof.

■

IIb Proof of Theorem 2

Note that a T_0-restricted finite Fourier Transform for a signal $z(t)$ is exactly identical to the full Fourier Transform of a T_0-restricted time-limited signal $z_{T_0}(t) = z(t)\mathbb{I}\{|t| < T_0\}$. Therefore, all the exposition in section 4.1 in the main manuscript still hold. In particular, frequency domain representation for aggregated data still follows equation 17.

Proof: We require a few modifications to our lemmata to derive the proof of Theorem 2. In the subsequent analysis, all Fourier Transforms should be assumed to be finite Fourier Transforms, but we omit the T superscript for notational succinctness.[2] We also assume that for every $\omega \in \Omega$ below, we have $|sin(\omega)| > \tau$ for some $\tau > 0$. This will not affect our algorithm because for small enough τ, as long as ω_0 is small enough in comparison to $\frac{2\pi}{T}$, the probability of sampling ω which violates this assumption is vanishingly small. In particular, this will be true for $\omega_0 \ll \omega_s/2$ where $\omega_s = \frac{2\pi}{T_s}$ with $T_s = \max\{T_y, T_1, T_2, \cdots T_d\}$.

Denote the reconstructed Fourier Transforms as

$$\widehat{X}_i(\omega) = \frac{\overline{X}_i(\omega)}{U(\omega)},\ \widehat{Y}(\omega) = \frac{\overline{Y}(\omega)}{U(\omega)}$$

Let $\omega_y = \frac{2\pi}{T_y}$ and $\omega_i = \frac{2\pi}{T_i}$. We have

$$\widehat{X}_i(\omega) = X_i(\omega) + \Lambda_{X_i}(\omega|\omega_i) \tag{13}$$
$$\widehat{Y}(\omega) = Y(\omega) + \Lambda_Y(\omega|\omega_y) \tag{14}$$

where, using the notation of section 4.1 of the main manuscript,

$$\Lambda_{X_i}(\omega|\omega_i) = \frac{1}{T_i}\sum_{k\in\mathbb{Z}\setminus\{0\}} X_i(\omega - k\omega_i)\frac{U(\omega-k\omega_i)}{U(\omega)} \tag{15}$$
$$\Lambda_Y(\omega|\omega_y) = \frac{1}{T_y}\sum_{k\in\mathbb{Z}\setminus\{0\}} Y(\omega - k\omega_y)\frac{U(\omega-k\omega_y)}{U(\omega)} \tag{16}$$

Let $\widehat{\mathbf{x}}(t), \widehat{y}(t)$ and $\lambda_i(t), \lambda_y(t)$ be the corresponding time domain signals. Use the following notation

$$\varepsilon_\beta(t) = \mathbf{x}(t)\boldsymbol{\beta} - y(t) \tag{17}$$
$$\hat{\varepsilon}_\beta(t) = \widehat{\mathbf{x}}(t)\boldsymbol{\beta} - \widehat{y}(t) \tag{18}$$
$$\varepsilon_{\lambda,\beta}(t) = \lambda_x(t)\boldsymbol{\beta} - \lambda_y(t) \tag{19}$$

Clearly, $\hat{\varepsilon}_\beta(t) = \varepsilon_\beta(t) + \varepsilon_{\lambda,\beta}(t)$. Denote the corresponding power spectral densities as $\hat{P}_{\hat{\varepsilon}_\beta}, P_{\varepsilon_\beta}, P_{\varepsilon_{\lambda,\beta}}$. We now show the following result

---

Lemma 4. *Suppose the power spectral densities of* $\mathbf{x}(t), y(t)$ *are finite for every* $\omega \in (-\omega_0, \omega_0)$*, and decay rapidly at a sub-Gaussian rate* $e^{-O((\omega-\omega_0)^2)}$ *beyond* $|\omega| > \omega_0$.

Then, we have, for any $\omega \in (-\omega_0, \omega_0)$

$$\hat{P}_{\hat{\varepsilon}_\beta}(\omega) - e^{-O((\omega_s-2\omega_0)^2)} \le P_{\varepsilon_\beta}(\omega) \le \hat{P}_{\hat{\varepsilon}_\beta}(\omega) + e^{-O((\omega_s-2\omega_0)^2)} \tag{20}$$

where $\omega_s = \frac{2\pi}{T_s}$ *with* $T_s = \max\{T_y, T_1, T_2, \cdots T_d\}$.

Proof. First, note that as a result of our assumptions, the power spectral densities $\hat{P}_{\hat{\varepsilon}_\beta}, P_{\varepsilon_\beta}, P_{\varepsilon_{\lambda,\beta}}$ are also finite for every $\omega \in (-\omega_0, \omega_0)$, and decays rapidly at a sub-Gaussian rate $e^{-O(\omega-\omega_0)^2}$ beyond $|\omega| > \omega_0$. Suppose $\hat{P}_{\hat{\varepsilon}_\beta}, P_{\varepsilon_\beta}, P_{\varepsilon_{\lambda,\beta}} < \gamma^2$ for some finite $\gamma > 0$.

The proof of this result requires two steps. First, suppose $g(t), h(t)$ are any two signals with corresponding (finite) power spectral densities P_g, P_h. Then, we have

$$P_{g+h} \le P_g + P_h + 2\sqrt{P_g P_h} \tag{21}$$

The proof of this is easy, and proceeds by simply expanding the expression for power spectral density and using standard results from real analysis and probability

[2] We also omit subscripts from the sinc function notation in the interest of succinctness, they will be clear from context.

theory.

$$\begin{aligned}P_{g+h} &= \lim_{T\uparrow\infty}\frac{1}{T}E\left[|G_T(\omega)+H_T(\omega)|^2\right]\\
&\leq \lim_{T\uparrow\infty}\frac{1}{T}E\left[|G_T(\omega)|^2+|H_T(\omega)|^2+2|G_T(\omega)H_T(\omega)|\right]\\
&\quad\text{(Triangle Inequality)}\\
&\leq \lim_{T\uparrow\infty}\frac{1}{T}\left[E|G_T(\omega)|^2+E|H_T(\omega)|^2\right]\\
&+\lim_{T\uparrow\infty}\frac{2}{T}\left[\sqrt{E[|G_T(\omega)H_T(\omega)|^2]}\right]\text{ (Jensen's Ineq.)}\\
&\leq \lim_{T\uparrow\infty}\frac{1}{T}E|G_T(\omega)|^2+\lim_{T\uparrow\infty}\frac{1}{T}E||H_T(\omega)|^2\\
&+2\sqrt{\lim_{T\uparrow\infty}\frac{1}{T}E[|G_T(\omega)|^2\lim_{T\uparrow\infty}\frac{1}{T}E|H_T(\omega)|^2]}\\
&\text{(Cauchy-Schwartz, limit theorems)}\\
&= P_g+P_h+2\sqrt{P_gP_h}\end{aligned}$$

Therefore, using this result, the definitions of $\hat{\varepsilon}_\beta(t), \varepsilon_\beta(t), \varepsilon_{\lambda,\beta}(t)$ and the fact that $P_{-z} = P_z$ for any signal z, we have,

$$\begin{aligned}\hat{P}_{\hat{\varepsilon}_\beta}(\omega) - &\left(P_{\varepsilon_{\lambda,\beta}}(\omega)+2\gamma\sqrt{P_{\varepsilon_{\lambda,\beta}}(\omega)}\right)\\
&\leq P_{\varepsilon_\beta}(\omega) \qquad (22)\\
&\leq \hat{P}_{\hat{\varepsilon}_\beta}(\omega)+\left(P_{\varepsilon_{\lambda,\beta}}(\omega)+2\gamma\sqrt{P_{\varepsilon_{\lambda,\beta}}(\omega)}\right)\end{aligned}$$

We can easily extend equation (21) to the following standard result. Suppose $z_i(t) : i = 1, 2, \cdots$ are an arbitrary set of signals. Then,

$$P_{\Sigma_i z_i} \leq \left(\sum_i \sqrt{P_{z_i}}\right)^2 \qquad (23)$$

This result works for infinite sums provided the right hand side exists finitely. The proof of this also proceeds by expanding the expression for power spectral density, and using standard limit theorems.

We shall use this to show that $P_{\varepsilon_{\lambda,\beta}}(\omega) \sim e^{-O(\omega_s-2\omega_0)^2}$. Define the following quantities

$$\begin{aligned}\Lambda_{X_i,k}(\omega|\omega_i) &= \frac{1}{T_i}X_i(\omega-k\omega_i)\frac{U(\omega-k\omega_i)}{U(\omega)}\\
\Lambda_{Y,k}(\omega|\omega_y) &= \frac{1}{T_y}Y(\omega-k\omega_y)\frac{U(\omega-k\omega_y)}{U(\omega)}\end{aligned}$$

Define $\lambda_{x_i,k}(t) = \mathcal{F}^{-1}\Lambda_{X_i,k}, \lambda_{y,k}(t) = \mathcal{F}^{-1}\Lambda_{Y,k}$. Clearly,

$$\Lambda_{X_i}(\omega|\omega_i) = \sum_{k\in\mathbb{Z}\setminus\{0\}}\Lambda_{X_i,k}(\omega|\omega_i) \qquad (24)$$

$$\Lambda_Y(\omega|\omega_y) = \sum_{k\in\mathbb{Z}\setminus\{0\}}\Lambda_{Y,k}(\omega|\omega_y) \qquad (25)$$

$$\lambda_i(t) = \sum_{k\in\mathbb{Z}\setminus\{0\}}\lambda_{x_i,k}(t) \qquad (26)$$

$$\lambda_y(t) = \sum_{k\in\mathbb{Z}\setminus\{0\}}\lambda_{y,k}(t) \qquad (27)$$

We note that for any signal $z(t)$, if $P_z(\omega) \sim e^{-O(\omega^2)}$ and $\tau(\omega)$ is a strictly bounded function of ω, then for $\lambda_z(t) = \mathcal{F}^{-1}Z(\omega)\tau(\omega)$, we have $P_\lambda(\omega) \sim e^{-O(\omega^2)}$.

By assumption, $P_{x_i}(\omega), P_y(\omega) \sim e^{-O(\omega-\omega_0)^2}$ and for the values of ω we use $\frac{U(\omega-k\omega_y)}{U(\omega)}$ is strictly bounded, therefore, we can show that $P_{\lambda_{x_i,k}}(\omega), P_{\lambda_{y,k}}(\omega) \sim e^{-O(\omega-\omega_0-k\omega_y)^2}$

We have, $\lambda_i(t) = \sum_k \lambda_{x_i,k}(t)$ and $\lambda_y(t) = \sum_k \lambda_{y,k}(t)$. Therefore, we have by equation 23,

$$\begin{aligned}P_{\lambda_i}(\omega) &= \left(\sum_{k\in\mathbb{Z}\setminus\{0\}}\sqrt{P_{\lambda_{x_i,k}}}\right)\\
&= 2\left(\sum_{k=1}^{\infty}\sqrt{P_{\lambda_{x_i,-k}}}\right)\text{ by symmetry around 0}\\
&= 2\sum_{k=1}^{\infty}e^{-O(k\omega_y+\omega-\omega_0)^2}\\
&\sim e^{-O(\omega_i-\omega_0+\omega)^2}\end{aligned}$$

Similarly, $P_{\lambda_y}(\omega) \sim e^{-O(\omega_y-\omega_0+\omega)^2}$. The final step uses standard approximation techniques exploiting the fact that $\sum_n f(n) \sim \Theta(\int_x f(x)dx)$ for bounded, finite, monotonic functions f, and noting that $e^{-O(k\omega_y+\omega-\omega_0)^2}$ has Gaussian decay in terms of k, and the area under Gaussian functions over a subset of the positive real line is given by the complementary error function $erfc(\cdot)$. We also use the fact[2] that the complementary error function has a Gaussian decay $erfc(x) \sim e^{-O(x^2)}$.

If $\omega_s = \min\{\omega_y, \omega_1, \omega_2, \cdots, \omega_d\}$, and for $\omega \in (-\omega_0, \omega_0)$, we have in terms of ω_s the fact that $e^{-O(\omega-\omega_0-\omega_y)^2} < e^{-O(\omega_s-2\omega_0)^2}$. For $\omega_s > 2\omega_0$, these approximations can be written more succinctly as $P_{\lambda_i}(\omega), P_{\lambda_y}(\omega) \sim e^{-O(\omega_s-2\omega_0)^2}$.

Finally, we note that by definition and using (23), we have $P_{\varepsilon_{\lambda,\beta}}(\omega) \leq (\beta_i \sum_{i=1}^d \sqrt{P_{\lambda_i}(\omega)} + \sqrt{P_{\lambda_y}(\omega)})^2$. For fixed d and since by assumption $|\boldsymbol{\beta}|$ is bounded, we have $P_{\varepsilon_{\lambda,\beta}}(\omega) \sim e^{-O(\omega_s-2\omega_0)^2}$ and therefore,

$\left(P_{\varepsilon_{\lambda,\beta}}(\omega) + 2\gamma\sqrt{P_{\varepsilon_{\lambda,\beta}}(\omega)}\right) \sim e^{-O(\omega_s - 2\omega_0)^2}$. This completes the proof for Lemma 4.

□

The final piece of the proof is to approximate $E\|\widehat{\mathbf{X}}_{T_0}(\omega) - \widehat{Y}_{T_0}(\omega)\|^2$. By assumption, the individual processes at each location is strictly sub-Gaussian [3, 4]. Simply put, this means that for each signal $z(t)$ at each time t, the logarithm of the moment generating function is quadratically bounded

$$\forall b > 0, \ln E[e^{b(z(t)-\mu)}] < \frac{b^2\sigma^2}{2}$$

for some constant σ, where $\mu = E[z(t)]$.

Since by assumption our random processes are bounded and almost surely finite, it can be shown by using results from calculus and probability theory that finite aggregation and Finite Fourier Transforms preserve sub-Gaussian property being linear operations[3]. In particular, note that most of our Fourier Transform computations can be estimated by discrete sums using the DTFT-DFT dual relationship, and linear sums preserve the sub-Gaussian property.

Now, we have that by using Hoeffding's inequality on sub-Gaussian random variables[5, 6], we can show that for independent observations $\{(\widehat{\mathbf{X}}^j(\omega), \widehat{Y}^j(\omega)) : j = 1, 2, \cdots N\}$ from N locations, for any small ξ, we have with probability $1 - exp(-O(N^2\xi_3^2))$,

$$E\|\widehat{\mathbf{X}}_{T_0}(\omega)\boldsymbol{\beta} - \widehat{Y}_{T_0}(\omega)\|^2 - \xi \tag{28}$$

$$< \frac{1}{N}\sum_{j\in[N]} \|\widehat{X}^j_{T_0}(\omega)^\top\boldsymbol{\beta} - \widehat{Y}^j_{T_0}(\omega)\|^2 \tag{29}$$

$$< E\|\widehat{\mathbf{X}}_{T_0}(\omega)\boldsymbol{\beta} - \widehat{Y}_{T_0}(\omega)\|^2 + \xi \tag{30}$$

Choose ξ such that $\xi_3 = (1 + \frac{1}{T_0})\xi$. Theorem 2 now follows in a manner exactly identical to the proof of Theorem 1, with the addition of two extra steps that incorporates Lemma 4 and equation 28. ■

Finally we note that Theorem 2 is only one of many possible results that can be obtained for estimation using our techniques. In particular, usage of different assumptions on the data distribution, and different decay rates on the power spectral densities can be used to derive alternative guarantees.

The proofs for results in the multidimensional case are exactly identical, except for the size of the sampled frequency set $|\Omega| = D$. As mentioned in the main manuscript, D can grow exponentially in the ambient dimensionality p of the interaction space $\mathbb{R}^p$. This is because the sampled frequencies are expected to cover a certain volume, and volume grows exponentially with dimensionality. However, in most real life cases, p will be very small (for example $p \leq 4$ for spatio-temporal applications), hence the increase in required size is in and of itself no major impediment in application of our algorithmic framework.

[3] An easy way to prove it, for example, would be to represent integration as the limit of a Riemann sum using definition from first principles, and to use the bounded convergence theorem and continuity of the exponentiation operator with standard limit theorems

III Discussion: Decay Rates

Throughout this manuscript, we assume that the power spectral density and autocovariance function for every signal of interest exists finitely for each ω. We further assume that the autocovariance function decays rapidly with lag for all processes involved in our analysis. In essence this means that the value of the time series at any given point is highly correlated with values at points close to it in time, but the correlation decreases rapidly with values farther away in time.

In particular, we assume that $\rho_{(\cdot)}(\cdot)$ is a Schwartz function [7], that is $\rho(\cdot)$ and all its derivatives decay at least as fast as any inverse polynomial. That is, $\forall \alpha, \beta \in \mathbb{Z}^n_+$ we have

$$|\zeta^\alpha \frac{\partial^\beta \rho(\zeta)}{\partial \zeta^n}| \to 0 \quad as \quad |\zeta| \to \infty$$

Examples of Schwartz functions are exponential functions like $e^{-a\zeta^2}$ for $a > 0$, or any polynomial $\wp(\zeta)$ multiplied with an exponential function like $\wp(\zeta)e^{-a\zeta^2}$, or any smooth domain-restricted function $f(\zeta)$ which is 0 outside of a bounded compact subset $\zeta \in \Im \subset \mathbb{R}^n$ (e.g. time limited signals).

A key property of Schwartz functions is that the Fourier Transform of a Schwartz function is itself a Schwartz function [8, 9]. Therefore, if we assume that the covariance functions $\rho_{(\cdot)}(\tau)$ decays rapidly with τ for each of our signals, then their corresponding power spectral densities $P_{(\cdot)}(\omega)$ will decay rapidly with ω, since $P = \mathcal{F}\rho$. Therefore, most of the power for our signals will be concentrated around $\omega = 0$.

As seen earlier, the decay rates of the power spectral density and autocovariance function complement each other- e.g., if ρ exhibits a Gaussian decay, then P_{ε_β} also exhibits a Gaussian decay. Similarly, if ρ exhibits power law or Lorentzian decay, then P_{ε_β} exhibits exponential decay. The exact decay rates involved will vary on a case to case basis, but in essence, this means that we only need to care about a small set of frequencies around 0 to describe the signal up to a reasonable approximation.

We note that unlike traditional signal processing applications, we do not consider a flat power spectral density (e.g. white noise) for our noise process. This is because traditional signal processing applications assume band-limited signals of interest. Properties of the noise process outside the band are irrelevant since outputs are going to be filtered regardless, and analysis only needs to focus on effects of additive noise within the frequency band of interest. In our case, we can make no such assumption–signals need not be bandlimited and therefore we have to consider effects of noise throught the entire spectrum[4].

References

[1] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

[2] Seok-Ho Chang, Pamela C Cosman, and Laurence B Milstein. Chernoff-type bounds for the gaussian error function. *IEEE Transactions on Communications*, 59(11):2939–2944, 2011.

[3] Valeriĭ Buldygin. *Metric characterization of random variables and random processes.*

[4] Shahar Mendelson. Discrepancy, chaining and subgaussian processes. *The Annals of Probability*, pages 985–1026, 2011.

[5] Panayiotis G Georgiou and Chris Kyriakakis. Robust maximum likelihood source localization: the case for sub-gaussian versus gaussian. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(4):1470–1480, 2006.

[6] Daniel Hsu, Sham M Kakade, and Tong Zhang. A tail inequality for quadratic forms of subgaussian random vectors. *Electron. Commun. Probab*, 17(52):1–6, 2012.

[7] T TerzioĞglu. On schwartz spaces. *Mathematische Annalen*, 182(3):236–242, 1969.

[8] Karlheinz Gröchenig and Georg Zimmermann. Hardy's theorem and the short-time fourier transform of schwartz functions. *Journal of the London Mathematical Society*, 63(1):205–214, 2001.

[9] Robert S Strichartz. *A guide to distribution theory and Fourier transforms.* World Scientific, 2003.

[4] Note that a true white noise process is unrealistic because it implies infinite variance for the noise process which renders any attempt at parameter learning futile.