Multi-kernel based nonlinear functional connectivity models

A THESIS
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Vasileios Georgios Karanikolas

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

Georgios B. Giannakis, Advisor

December, 2016

Acknowledgments

First and foremost, I would like to thank my advisor Prof. Georgios B. Giannakis for his guidance and encouragement as well as the significant amount time of time he has put towards improving my skills as a researcher.

I am also grateful to Professors Jarvis D. Haupt and Kendrick Kay for insightful discussions and for serving on the committee of this thesis.

Special thanks go to Professors Richard M. Leahy and Konstantinos Slavakis for their input as collaborators.

At this point, I would also like to express my gratitude to Prof. Olaf Sporns for providing the real data used in this thesis.

Finally, I am very grateful to my family and friends for their support.

This work was supported by the following grants: NSF 1500713, 1514056, and NIH 1R01GM104975-01.

Abstract

Functional connectivity measures, such as partial correlation (PC) and Granger causality, play a key role in identifying interactions among brain regions from functional magnetic resonance imaging (fMRI) time series. Motivated by the generally nonlinear mechanisms generating the blood-oxygen-level dependent signal, the present thesis introduces kernel-based nonlinear counterparts of partial correlation and partial Granger causality (PGC). The form of kernel-induced nonlinearity that "best" models the data is learned through a data-driven approach that optimally combines multiple kernels. Synthetically generated data based on a dynamic causal model are used to validate the proposed approaches in resting-state (RS) fMRI scenarios, highlighting the gains in edge presence and directionality detection performance when compared with the linear PC and existing PGC methods, respectively. Tests on real RS-fMRI data demonstrate that connectivity patterns revealed by linear and nonlinear models exhibit noticeable differences. In particular, the networks estimated by the proposed kernel-based PC approach capture known features of RS networks, while at the same time being more reflective of the underlying structural connectivity, as compared to linear PC networks.

Contents

Appendices

List of Figures

Chapter 1

Introduction

Functional (f)MRI is a powerful tool for estimating brain activity, and has provided great insights with regards to brain functionality [1]. Different from initial fMRI studies, many recent works treat the brain as a network [2], adopting tools and notions developed in the context of network science [3]. Key to obtaining the the topology of such networks from fMRI time series is an appropriate functional connectivity measure [4], that is able to sufficiently capture statistical dependencies between neuronal activities of distinct brain regions.

Typical examples of functional connectivity measures include (Pearson) correlation [4], and partial correlation (PC) coefficients [5]. However, (partial) correlation can only identify edge presence. When edge directionality is also desired, (linear) Granger causality and its variants, such as partial Granger causality (PGC), are popular approaches [6, 7].

Typically, PC-based topology identification of brain graphs entails an estimate of the inverse covariance matrix of the nodal time series obtained by maximizing a regularized version of a likelihood-based criterion. Regularization aims at promoting sparse graphs, with the elastic net [8], and the ℓ_1-norm [4, 9] being typical examples. The desired PC coefficients can then be obtained using entries of the estimated inverse covariance matrix [9].

PC aims at unveiling direct interactions between pairs of brain regions by adjusting for the influence of (all) other regions. One limitation of linear PC however, is that it can fully remove only linear influences exerted by other regions. Additional motivation

for developing nonlinear approaches is provided by observations suggesting that the blood-oxygen-level dependent (BOLD) response is a nonlinear function of the underlying neural activity [10].

The present contribution adopts a kernel-based nonlinear regression approach to estimate the PC coefficients as well as the PGC test statistics, on the premise that *nonlinear* estimators will offer an improved fit of the data as compared to *linear* ones. Motivation for this novel approach also stems from the fact that generally nonlinear models can capture nonlinear dependencies that linear models are unable to pick up, and in the case of PC, the former will be more successful than the latter in accounting for nonlinear influences.

The problem of choosing the kernel, upon which the performance of any kernel-based method is highly dependent, is tackled here using multi-kernel learning. This data-driven approach learns a combination of kernels taken from a preselected dictionary of kernel functions, with the goal of optimizing the fit to the data. Finally, the kernel-based variants of PC and PGC connectivity measures are developed under a common framework that is flexible enough to accommodate alternative loss functions and regularization terms.

Although linear Granger causality is more prevalent [7], a kernel-based variant thereof has been reported [11, 12]; see also [6] for a nonlinear version of PGC. Unlike the present thesis however, [6] does not follow (and thus benefit from) a reproducing kernel Hilbert space formulation, whereas [11, 12] do not deal with the critical issue of kernel selection. Moreover, neither of these approaches has been tested on "ground-truth" synthetics based on dynamic causal modeling (DCM) [4, 13].

At a broader level, kernel-based methods have been used in fMRI studies with examples including kernel Granger causality, and kernel canonical correlation analysis [14]. Multi-kernel learning techniques have also been applied in this context, with the focus however being on different tasks such as classification and feature selection [15, 16], as well as data fusion from heterogeneous sources [17].

The rest of this thesis is organized as follows. First, kernel-based PC (KPC) is formulated [18], and a nonlinear kernel-based variant of PGC is developed in Sec. 2.1, under a common framework. A suitable multi-kernel learning approach is the subject of Sec. 2.2

along with algorithms for computing the novel functional connectivity measures. Tests on DCM based synthetics, comparing the proposed approaches with existing ones are presented in Sec. 2.3.1, followed by tests on real RS-fMRI data, before concluding.

Chapter 2

Nonlinear Connectivity Models

2.1 Nonlinear connectivity models

2.1.1 Undirected graph topology identification

Consider an anatomical or data-driven parcellation of the brain into regions (represented by nodes), and let $\mathcal{V}$ denote the set of all such nodes. Per region $\nu \in \mathcal{V}$, a representative vector $\mathbf{x}_\nu := [x_\nu[1] \dots x_\nu[T]]^\top$ ($^\top$ stands for transposition), is obtained from the time series of voxels belonging to the same region.

The immediate and simple measure to identify connectivity between pairs of regions (i, j) is to evaluate the correlation among representatives $\mathbf{x}_i$ and $\mathbf{x}_j$; that is, with $\mathbf{1}$ denoting the all-one $T \times 1$ vector, and $\bar{\mathbf{x}}_i := T^{-1} \sum_{t=1}^{T} x_i[t]\mathbf{1}$, region i will be deemed connected with j depending on the value of

$$\rho_{ij} := \frac{(\mathbf{x}_i - \bar{\mathbf{x}}_i)^\top (\mathbf{x}_j - \bar{\mathbf{x}}_j)}{\|\mathbf{x}_i - \bar{\mathbf{x}}_i\|_2 \| \mathbf{x}_j - \bar{\mathbf{x}}_j\|_2}. \tag{2.1}$$

For a prescribed probability of false alarm, a threshold τ is determined to decide whether a non-directional edge is absent ($|\rho_{ij}| < \tau$) or not.

However, having $|\rho_{ij}| > \tau$ does not indicate whether the dependence of region i with region j is mediated or not. To highlight this distinction, consider a network of three nodes in cascade $i \to k \to j$ [4]. All three vectors $\mathbf{x}_i$, $\mathbf{x}_j$, $\mathbf{x}_k$ are mutually correlated, which erroneously assigns an edge connecting i with j, even though nodes i and j are only linked

through the mediating node k. To account for this mediation, one can regress out $\mathbf{x}_k$ from $\mathbf{x}_i$ and $\mathbf{x}_j$ and thus remove the correlation (edge) connecting regions i and j.

To introduce the measure identifying unmediated (i,j) dependencies, consider estimated vectors $\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_j$ at nodes $i,j \in \mathcal{V}$, found using data $\{\mathbf{x}_k \mid k \in \mathcal{S}\}$, where $\mathcal{S}$ is a subset of nodes satisfying $\mathcal{S} \subseteq \mathcal{V}\backslash\{i,j\}$, with $\backslash$ representing set difference. Letting $\tilde{\mathbf{x}}_i := \mathbf{x}_i - \hat{\mathbf{x}}_i$ and likewise for $\tilde{\mathbf{x}}_j$, the sample PC coefficient of $\mathbf{x}_i, \mathbf{x}_j$ with respect to $\{\mathbf{x}_k\}_{k\in\mathcal{S}}$ is given by

$$\hat{\rho}_{ij|\mathcal{S}} := \frac{(\tilde{\mathbf{x}}_i - \bar{\tilde{\mathbf{x}}}_i)^\top (\tilde{\mathbf{x}}_j - \bar{\tilde{\mathbf{x}}}_j)}{\|\tilde{\mathbf{x}}_i - \bar{\tilde{\mathbf{x}}}_i\|_2 \| \tilde{\mathbf{x}}_j - \bar{\tilde{\mathbf{x}}}_j\|_2} \tag{2.2}$$

where $\bar{\tilde{\mathbf{x}}}_i := T^{-1}\sum_{t=1}^T \tilde{x}_i[t]\mathbf{1}$. Assessing whether an edge is present between a pair of nodes (i,j) requires performing a hypothesis test on the statistic $\hat{\rho}_{ij|\mathcal{S}}$ (or a function thereof), as it will be detailed in Sec. 2.1.3. Notice that both measures (2.1) and (2.2) are symmetric in (i,j); and therefore, they cannot be used to assess directionality.

Kernel-based nonlinear PC measures

Typically, the estimate $\hat{\mathbf{x}}_i$ is a linear function of $\{\mathbf{x}_k \mid k \in \mathcal{S}\}$; see e.g., [4, 9]. The premise behind our approach however, is that using a nonlinear function instead will produce at least as reliable estimators, while at the same time rendering more general connectivity models attainable. In order to model the aforementioned nonlinear functions, we will rely on a reproducing kernel Hilbert space (RKHS) formulation [19].

Let indices $\{n_{1\backslash ij}, \ldots, n_{|V|-2\backslash ij}\}$ enumerate nodes in $\mathcal{S} := \mathcal{V}\backslash\{i,j\}$, and $\boldsymbol{\chi}_{\backslash ij}[t] := [x_{n_{1\backslash ij}}[t], \ldots, x_{n_{|V|-2\backslash ij}}[t]]^\top$ denote the observations at all nodes, except for i and j, at time slot t. The time series per node i will be modeled as

$$x_i[t] = f_i(\boldsymbol{\chi}_{\backslash ij}[t]) + \epsilon_i[t] \tag{2.3}$$

where $\epsilon_i[t]$ captures noise and unmodeled effects, while f_i is a nonlinear function from the space of functions given by

$$\mathcal{H} := \{f : f(\boldsymbol{\chi}[t]) = \sum_{\tau=1}^{\infty} \beta_\tau \kappa(\boldsymbol{\chi}[t], \boldsymbol{\chi}[\tau])\} \tag{2.4}$$

with preselected basis functions κ (a.k.a. kernels) that measure "similarity" of $\boldsymbol{\chi}[t]$ with $\boldsymbol{\chi}[\tau]$. Examples of widely used kernel functions κ, which define uniquely their associated

RKHSs $\mathcal{H}$, include the linear kernel $\kappa_L(\boldsymbol{\chi}_1, \boldsymbol{\chi}_2) := \boldsymbol{\chi}_1^\top \boldsymbol{\chi}_2$, and the Gaussian kernel [19]

$$\kappa_{\mathrm{G}}(\boldsymbol{\chi}_1, \boldsymbol{\chi}_2) := e^{\frac{-\|\boldsymbol{\chi}_1 - \boldsymbol{\chi}_2\|_2^2}{2\sigma^2}}. \tag{2.5}$$

Kernels such as κ_{G} are reproducing in the sense that their inner product satisfies $\langle \kappa(\boldsymbol{\chi}[t], \boldsymbol{\chi}[\tau]), \kappa(\boldsymbol{\chi}[t], \boldsymbol{\chi}[\tau'])\rangle = \kappa(\boldsymbol{\chi}[\tau], \boldsymbol{\chi}[\tau'])$. After using the linearity of inner products, the reproducing property implies that for $f, f' \in \mathcal{H}$ it holds that $\langle f, f' \rangle_{\mathcal{H}} = \sum_\tau \sum_{\tau'} \beta_\tau \beta'_{\tau'} \kappa(\boldsymbol{\chi}[\tau], \boldsymbol{\chi}[\tau'])$, which yields the $\mathcal{H}$-norm as $\|f\|_{\mathcal{H}}^2 = \sum_\tau \sum_{\tau'} \beta_\tau \beta_{\tau'} \kappa(\boldsymbol{\chi}[\tau], \boldsymbol{\chi}[\tau'])$.

Given T samples $\{x_i[t]\}_{t=1}^T$ per node i, RKHS based nonparametric regression seeks an interpolating function that optimally fits the data while adhering to the norm of $\mathcal{H}$; that is,

$$\hat{f}_i = \underset{f \in \mathcal{H}}{\arg\min} \sum_{\tau=1}^{T} (x_i[\tau] - f(\boldsymbol{\chi}_{\backslash ij}[\tau]))^2 + \lambda \|f\|_{\mathcal{H}}^2 \tag{2.6}$$

where λ is a regularization parameter. For the functional optimization task in (2.6), the representer theorem [20] asserts that the optimal solution is given by a linear superposition of kernels located[1] at the T samples

$$\hat{f}_i(\boldsymbol{\chi}_{\backslash ij}[t]) = \sum_{\tau=1}^{T} \beta_{i\tau}\, \kappa(\boldsymbol{\chi}_{\backslash ij}[t], \boldsymbol{\chi}_{\backslash ij}[\tau]). \tag{2.7}$$

Letting the kernel matrix $\mathbf{K}_{\backslash ij}$ be formed with (t, τ) entries $[\mathbf{K}_{\backslash ij}]_{t\tau} = \kappa(\boldsymbol{\chi}_{\backslash ij}[t], \boldsymbol{\chi}_{\backslash ij}[\tau])$, and plugging (2.7) into (2.6), the functional minimization problem boils down to estimating the vector $\boldsymbol{\beta}_i := [\beta_{i_1}, \ldots, \beta_{i_T}]^\top$, as

$$\hat{\boldsymbol{\beta}}_i = \underset{\boldsymbol{\beta}_i \in \mathbb{R}^T}{\arg\min} \|\mathbf{x}_i - \mathbf{K}_{\backslash ij} \boldsymbol{\beta}_i\|^2 + \lambda \boldsymbol{\beta}_i{}^\top \mathbf{K}_{\backslash ij} \boldsymbol{\beta}_i \tag{2.8}$$

where we used that $\|f\|_{\mathcal{H}}^2 = \boldsymbol{\beta}_i{}^\top \mathbf{K}_{\backslash ij} \boldsymbol{\beta}_i$. The ridge regression task in (2.8) admits the following closed-form solution

$$\hat{\boldsymbol{\beta}}_i = (\mathbf{K}_{\backslash ij} + \lambda \mathbf{I})^{-1} \mathbf{x}_i. \tag{2.9}$$

[1]From vantage point, kernels can be viewed as generalizations of the *sinc* function in the reconstruction formula of bandlimited continuous waveforms from their T samples.

Based on (2.3), (2.7), and (2.9), the $\hat{\mathbf{x}}_i$ kernel-based estimate can be expressed as $\hat{\mathbf{x}}_i = [\hat{f}_i(\boldsymbol{\chi}_{\setminus ij}[1]) \dots \hat{f}_i(\boldsymbol{\chi}_{\setminus ij}[T])]^\top = \mathbf{K}_{\setminus ij}\hat{\boldsymbol{\beta}}_i$, that is

$$\hat{\mathbf{x}}_i = \mathbf{K}_{\setminus ij}(\mathbf{K}_{\setminus ij} + \lambda \mathbf{I})^{-1}\mathbf{x}_i. \tag{2.10}$$

Using the latter in forming $\tilde{\mathbf{x}}_i$ (and likewise for $\tilde{\mathbf{x}}_j$) in (2.2), leads to a kernel-based PC measure.

Alternative loss functions and regularizers

At this point, it is worth mentioning that replacing the square loss of the previous section with other loss functions, does not alter the general form of the solution (2.7). As an example, the so-termed ε-insensitive loss, defined as

$$\mathcal{L}_\varepsilon(x) := \begin{cases} 0, & |x| \leq \varepsilon \\ |x| - \varepsilon, & \text{otherwise} \end{cases}$$

can replace the square loss in (2.6) to obtain

$$\hat{f}_i = \underset{f \in \mathcal{H}}{\arg\min} \sum_{\tau=1}^{T} \mathcal{L}_\varepsilon(x_i[\tau] - f(\boldsymbol{\chi}_{\setminus ij}[\tau])) + \lambda \|f\|^2_{\mathcal{H}}. \tag{2.11}$$

It is known that $\mathcal{L}_\varepsilon$ is particularly well suited if the noise in (2.3) is assumed to (approximately) follow a uniform distribution [21]. Finally, off-the-shelf convex optimization packages, such as CVX [22,23] can be used to solve the resulting optimization problem, once the solution form (2.7) is plugged into (2.11). Additional choices of loss functions include the Huber and the ℓ_1 loss, both of which are cost functions that can cope with outliers [24].

Besides loss functions, the representer theorem allows for more general regularization terms that are increasing functions of $\|f\|^2_{\mathcal{H}}$, such as $\|f\|_{\mathcal{H}}$ for example; see also [20].

2.1.2 Directed graph topology identification

As mentioned after (2.2), one limitation of partial correlation is that it does not provide information about directionality. For a pair of nodes (i, j) one approach to assess whether

j causes i is to test whether including previous values of $\{x_j[\tau]\}_{\tau<t}$ in the regressors for estimating $x_i[t]$ increases the accuracy of $\hat{x}_i[t]$.

Such an approach is offered by PGC which boils down to estimating the residuals $\{\epsilon_{i|V\setminus j}[t]\}_{t=d+1}^{T}$, $\{\epsilon_{i|V}[t]\}_{t=d+1}^{T}$ in the following two d-th order linear regression models [25]

$$x_i[t] = \bar{\boldsymbol{\chi}}_{\setminus ij}^{\top}[t]\boldsymbol{\gamma}_i + \epsilon_{i|V\setminus j}[t], \quad t = d+1,\ldots,T \tag{2.12a}$$

$$x_i[t] = \bar{\boldsymbol{\chi}}_{\setminus ij}^{\prime\top}[t]\boldsymbol{\delta}_i + \epsilon_{i|V}[t], \quad t = d+1,\ldots,T \tag{2.12b}$$

$$\bar{\boldsymbol{\chi}}_{\setminus ij}[t] := [\boldsymbol{\chi}_{\setminus ij}^{\top}[t],\ldots,\boldsymbol{\chi}_{\setminus ij}^{\top}[t-d], x_i[t-1],\ldots,x_i[t-d]]^{\top} \tag{2.12c}$$

$$\bar{\boldsymbol{\chi}}_{\setminus ij}^{\prime}[t] := [\bar{\boldsymbol{\chi}}_{\setminus ij}^{\top}[t], x_j[t-1],\ldots,x_j[t-d]]^{\top} \tag{2.12d}$$

where $\bar{\boldsymbol{\chi}}_{\setminus ij}^{\prime}[t]$ augments $\bar{\boldsymbol{\chi}}_{\setminus ij}[t]$ with the d past observations $\{x_j[\tau]\}_{\tau=t-d}^{t-1}$. If model (2.12b) is deemed statistically more valid than (2.12a), then a (directed) edge from j to i is declared to be present in the estimated graph. This is the basic idea behind partial Granger causality (PGC), that offers also robustness to exogenous inputs and latent variables as compared to the "ordinary" GC [6, 26].

To assess whether (2.12a) or (2.12b) is in effect, a hypothesis test is required. Specifically, one relies on the test statistic

$$F_{ij} := \frac{\mathrm{var}(\epsilon_{i|V\setminus j})}{\mathrm{var}(\epsilon_{i|V})} \tag{2.13}$$

which captures the ratio of the residual variances in (2.12a) and (2.12b), to perform the following hypothesis test

$$\mathrm{H}_0 : F_{ij} \leq 1; \quad \mathrm{H}_1 : F_{ij} > 1\,.$$

According to H_1, $\{x_j[t]\}$ "partial Granger causes" $\{x_i[t]\}$, whereas according to H_0 it does not.

Kernel-based PGC measures

Using the framework outlined in Sec. 2.1.1 we can now introduce the novel kernel-based (K)PGC. First, the regression models (2.12a), (2.12b) will be replaced by their kernel-based

counterparts, namely

$$x_i[t] = f_{i|V\setminus j}(\bar{\boldsymbol{\chi}}_{\setminus ij}[t]) + \epsilon_{i|V\setminus j}[t] \tag{2.14a}$$

$$x_i[t] = f_{i|V}(\bar{\boldsymbol{\chi}}'_{\setminus ij}[t]) + \epsilon_{i|V}[t] \tag{2.14b}$$

where $f_{i|V\setminus j}$, $f_{i|V} \in \mathcal{H}$. Recognizing the resemblance of (2.14a), (2.14b) with (2.3), we formulate two optimization problems similar to (2.6). Invoking once again the representer theorem, and plugging the solution forms for $f_{i|V\setminus j}$, $f_{i|V}$ in (2.14a), (2.14b), the regression models obtained are

$$x_i[t] = \sum_{\tau=d+1}^{T} \zeta_{i\tau}\, \kappa(\bar{\boldsymbol{\chi}}_{\setminus ij}[t], \bar{\boldsymbol{\chi}}_{\setminus ij}[\tau]) + \epsilon_{i|V\setminus j}[t] \tag{2.15a}$$

$$x_i[t] = \sum_{\tau=d+1}^{T} \eta_{i\tau}\, \kappa(\bar{\boldsymbol{\chi}}'_{\setminus ij}[t], \bar{\boldsymbol{\chi}}'_{\setminus ij}[\tau]) + \epsilon_{i|V}[t]. \tag{2.15b}$$

Accordingly, the associated optimization problems can be expressed as [cf. (2.8)]

$$\begin{aligned}\hat{\boldsymbol{\zeta}}_i &= \arg\min_{\boldsymbol{\zeta}_i} \|\mathbf{x}_i^{(d)} - \mathbf{K}_{i|V\setminus j}\boldsymbol{\zeta}_i\|_2^2 + \lambda {\boldsymbol{\zeta}_i}^\top \mathbf{K}_{i|V\setminus j}\boldsymbol{\zeta}_i \qquad (2.16a)\\ &= (\mathbf{K}_{i|V\setminus j} + \lambda\mathbf{I})^{-1}\mathbf{x}_i^{(d)}\\ \hat{\boldsymbol{\eta}}_i &= \arg\min_{\boldsymbol{\eta}_i} \|\mathbf{x}_i^{(d)} - \mathbf{K}_{i|V}\boldsymbol{\eta}_i\|_2^2 + \lambda {\boldsymbol{\eta}_i}^\top \mathbf{K}_{i|V}\boldsymbol{\eta}_i \qquad (2.16b)\\ &= (\mathbf{K}_{i|V} + \lambda\mathbf{I})^{-1}\mathbf{x}_i^{(d)}\end{aligned}$$

where $\mathbf{x}_i^{(d)} := [x_i[d+1], \ldots, x_i[T]]^\top$, $\boldsymbol{\zeta}_i := [\zeta_{i(d+1)}, \ldots, \zeta_{iT}]^\top$ and $\boldsymbol{\eta}_i := [\eta_{i(d+1)}, \ldots, \eta_{iT}]^\top$. Moreover, the (t,τ) entries of the kernel matrices in (2.16a) and (2.16b) are given by $[\mathbf{K}_{i|V\setminus j}]_{t\tau} = \kappa(\bar{\boldsymbol{\chi}}_{\setminus ij}[t], \bar{\boldsymbol{\chi}}_{\setminus ij}[\tau])$ and $[\mathbf{K}_{i|V}]_{t\tau} = \kappa(\bar{\boldsymbol{\chi}}'_{\setminus ij}[t],\ \bar{\boldsymbol{\chi}}'_{\setminus ij}[\tau])$, respectively. By plugging the estimated coefficients $\hat{\boldsymbol{\zeta}}_i$, $\hat{\boldsymbol{\eta}}_i$ into (2.15a) and (2.15b) one obtains $\{\epsilon_{i|V\setminus j}[t]\}_{t=d+1}^T$ and $\{\epsilon_{i|V}[t]\}_{t=d+1}^T$. Finally, the sample variances of $\epsilon_{i|V\setminus j}$ and $\epsilon_{i|V}$ are used to calculate the corresponding test statistic F_{ij}.

At this point, it is worth mentioning that application of Granger causality to fMRI data has been met with some skepticism, e.g. due to variability of the hemodynamic lags across the brain [7, 27]. The goal of the proposed approach however, is to obtain an improved estimate of the PGC test statistic, by accounting for nonlinearities in the regression models

involved; hence, if variability challenges the fMRI data, our approach remains operational even with EEG data.

2.1.3 Edge inference

In this subsection, the novel kernel-based (K)PC and (K)PGC measures will be employed to adapt statistical tests needed to decide the presence (or absence) and directionality of edges in the sought functional connectivity network.

Edge presence or absence

With $\mathcal{E} := \{(i,j) \in \mathcal{V} \times \mathcal{V} \mid \rho_{ij|\mathcal{S}} \neq 0\}$, the hypothesis testing problem for the potential presence of edge (i,j) is

$$\mathrm{H}_0 : \rho_{ij|\mathcal{S}} = 0; \quad \mathrm{H}_1 : \rho_{ij|\mathcal{S}} \neq 0\,. \tag{2.17}$$

Problem (2.17) will be solved using as test statistic the Fisher-z transformation [9], which in our context is given by

$$z^{(T)}_{ij|\mathcal{S}} := \frac{1}{2} \ln \left(\frac{1 + \hat{\rho}_{ij|\mathcal{S}}}{1 - \hat{\rho}_{ij|\mathcal{S}}} \right) \tag{2.18}$$

where the superscript (T) stresses the dependence of $\hat{\rho}_{ij|\mathcal{S}}$ on T. As $T \to \infty$, $z^{(T)}_{ij|\mathcal{S}}$ is zero-mean Gaussian with variance $\sigma^2_{ij|\mathcal{S}} = 1/[T - (|\mathcal{V}| - 2) - 3]\ \forall (i,j)$, i.e. $z_{ij|\mathcal{S}} \sim \mathcal{N}(0, \sigma^2_{ij|\mathcal{S}})$ [9], and it is preferred over $\hat{\rho}_{ij|\mathcal{S}}$ because it approaches normality with T smaller than PCs [28].

In the base case $|\mathcal{V}| = 2$, one hypothesis test (2.17) is sufficient, since there is only one potential edge. Given that (2.17) is a two-sided hypothesis test [29], deciding which hypothesis is in effect amounts to comparing $|z^{(T)}_{ij|\mathcal{S}}|$, instead of $z^{(T)}_{ij|\mathcal{S}}$, with a threshold; that is,

$$|z^{(T)}_{ij|\mathcal{S}}| \underset{\mathrm{H}_0}{\overset{\mathrm{H}_1}{\gtrless}} \tau. \tag{2.19}$$

Once a probability of false alarms (rejecting H_0 when it is in effect) P_{FA} is fixed, the appropriate threshold τ is the one satisfying $P_{FA} = 2(1 - \Phi_Z(\tau))$, where Φ_Z denotes the cumulative distribution function of $Z \sim \mathcal{N}(0, \sigma^2_{ij|\mathcal{S}})$.

Brain networks however, in general have $|\mathcal{V}| \geq 2$ nodes and therefore there are multiple tests to be performed, namely $\binom{|\mathcal{V}|}{2}$ for (K)PC to detect edge presence, and twice as many for (K)PGC to decide directionality. When multiple tests are jointly performed, one option is to control for the probability of a single false alarm (FA) in all tests, by adjusting the (common) threshold τ (e.g. using the Bonferroni correction [9]). Such an adjustment however, results in a large number of false rejections of H_1. One way to alleviate this, is to employ a different criterion, such as the false discovery rate $\text{FDR} := \mathbb{E}[N_{FA}/N_D]$, which is defined as the expected ratio of the number of false alarms (N_{FA}) over the number of discoveries[2] (N_D) [30, 31].

As an illustrative example, briefly describing an approach to control the FDR [31] in the PC case, is in order. Let $\mathcal{T}_{ij}$ denote the test for the potential edge (i, j) and $p_{ij} := 2(1 - \Phi_Z(z_{ij|\mathcal{S}}))$ the corresponding p-value. Moreover, let $p_{(1)}, \ldots, p_{(\binom{|\mathcal{V}|}{2})}$ denote the sorted set of p-values, α the desired maximum FDR level, and I the largest integer for which $p_{(I)} \leq (I/\binom{|\mathcal{V}|}{2})\alpha$. We then decide that H_1 is in effect in $\mathcal{T}_{(1)}, \ldots, \mathcal{T}_{(I)}$, whereas H_0 holds for the rest of the tests.

Edge directionality

Since the distribution of the test statistic F_{ij} under H_0 is unknown, the stationary bootstrap scheme of [32] can be employed to obtain an estimate thereof. The (mean) block length for the stationary bootstrap is chosen using the data-driven approach described in [33]. For each pair of nodes, say (i, j), two hypothesis tests are performed, to decide whether $\{x_j[t]\}$ "partial Granger causes" $\{x_i[t]\}$, and vice versa.

2.2 Multi-kernel based learning

The choice of the kernel κ in the RKHS formulation (2.4) affects critically the predictors $\{\hat{\mathbf{x}}_i\}$, and thus the edge inference and directionality assessment performance. A systematic selection is to rely on the data and choose a linear combination of kernels from a preselected dictionary [34].

[2] A discovery is defined as a rejection of H_0 regardless of it being true or not.

There are several ways of combining kernels, but here we will consider nonnegative superpositions of the basis kernels. Specifically, given P preselected (reproducing) kernel functions $\{\kappa_p\}_{p=1}^P$, the sought kernel κ is obtained via the linear combination $\kappa := \sum_{p=1}^P \theta_p \kappa_p$, where $\boldsymbol{\theta} := [\theta_1, \ldots, \theta_P]^\top \succeq \mathbf{0}$. Since $\boldsymbol{\theta} \succeq \mathbf{0}$, as long as $\kappa_1, \ldots, \kappa_P$ are reproducing, so will be κ [19]. In terms of kernel matrices, the previous combination gives rise to the multi-kernel matrix $\mathbf{K}_{\setminus ij} = \sum_{p=1}^P \theta_p \mathbf{K}_{p\setminus ij}$.

As a first step towards specifying $\{\theta_p\}_{p=1}^P$, consider the following problem

$$\min_{\boldsymbol{\beta}_i \in \mathbb{R}^T} \|(1/\sqrt{\lambda})\mathbf{x}_i - \sqrt{\lambda}\boldsymbol{\beta}_i\|^2 + \boldsymbol{\beta}_i{}^\top \mathbf{K}_{\setminus ij} \boldsymbol{\beta}_i. \tag{2.20}$$

By equating to zero the gradient w.r.t. $\boldsymbol{\beta}_i$, the optimum solution to (2.20) is also given by (2.9). Plugging now the multi-kernel $\mathbf{K}_{\setminus ij}$ in (2.20), dropping irrelevant terms, and minimizing the resulting cost w.r.t. $\boldsymbol{\theta}$, we arrive at the following optimization task

$$\min_{\boldsymbol{\theta} \in \Theta} \min_{\boldsymbol{\beta}_i \in \mathbb{R}^T} \lambda \boldsymbol{\beta}_i^\top \boldsymbol{\beta}_i - 2\boldsymbol{\beta}_i^\top \mathbf{x}_i + \sum_{p=1}^P \theta_p \boldsymbol{\beta}_i^\top \mathbf{K}_{p\setminus ij} \boldsymbol{\beta}_i. \tag{2.21a}$$

After interchanging the minimization tasks, (2.21a) reduces to

$$\min_{\boldsymbol{\beta}_i \in \mathbb{R}^T} \lambda \boldsymbol{\beta}_i^\top \boldsymbol{\beta}_i - 2\boldsymbol{\beta}_i^\top \mathbf{x}_i - \min_{\boldsymbol{\theta} \in \Theta} -\boldsymbol{\theta}^\top \mathbf{v} \tag{2.21b}$$

where $\mathbf{v} := [v_1, \ldots, v_P]^\top$ with $v_p := \boldsymbol{\beta}_i^\top \mathbf{K}_{p\setminus ij} \boldsymbol{\beta}_i$, and Θ is a set constraining the term $-\boldsymbol{\theta}^\top \mathbf{v}$ from becoming arbitrarily small. In particular, here, we adopt

$$\Theta := \{\boldsymbol{\theta} \in \mathbb{R}^P : \boldsymbol{\theta} \succeq \mathbf{0}, \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_2 \leq \Lambda\}, \tag{2.22}$$

for a pre-defined $\Lambda > 0$ and $\boldsymbol{\theta}_0 \in \mathbb{R}^P$, which amounts to applying an ℓ_2 regularizer to the weight vector $\boldsymbol{\theta}$.

For any $\boldsymbol{\beta}_i$, the optimal solution w.r.t. $\boldsymbol{\theta}$ is obtained from the Karush-Kuhn-Tucker (KKT) conditions on the convex problem $\min_{\boldsymbol{\theta} \in \Theta} -\boldsymbol{\theta}^\top \mathbf{v}$. This optimum is clearly attained at the boundary of Θ, and it is given by

$$\boldsymbol{\theta}^\star(\boldsymbol{\beta}_i) := \boldsymbol{\theta}_0 + \Lambda \mathbf{v}(\boldsymbol{\beta}_i)/\|\mathbf{v}(\boldsymbol{\beta}_i)\|_2. \tag{2.23a}$$

In order to find the optimal $\boldsymbol{\beta}_i$, it suffices to plug $\boldsymbol{\theta}^\star$ into (2.21b) and solve the resulting minimization problem. With the optimal solution being

$$\boldsymbol{\beta}_i^\star(\boldsymbol{\theta}^\star) := [\mathbf{K}_{\setminus ij}(\boldsymbol{\theta}^\star) + \lambda \mathbf{I}_T]^{-1}\mathbf{x}_i \tag{2.23b}$$

one observes that $\boldsymbol{\beta}_i^\star$ is a function of $\boldsymbol{\theta}^\star$ and vice versa. Adopting [35], this observation motivates the iterative Algorithm 1, which alternates between (2.23b) and (2.23a).

In the KPC case, the estimates $\{\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_j\}_{i,j\in\mathcal{V}}$ obtained via Alg. 1 are plugged into (2.2) to compute the wanted coefficients $\hat{\rho}_{ij|\mathcal{S}}$. A slight modification of Alg. 1 detailed in Appendix B, also accommodates the KPGC case.

At this point, it should be noted that alternative regularizers can also be used, an example being the ℓ_1 norm, which promotes sparsity in the multi-kernel combination [34].

2.3 Numerical tests

2.3.1 Synthetic data

Performance of the proposed approaches was tested using synthetic fMRI data based on the forward dynamic causal model (DCM), in a setup similar to that of [4]. Let $\psi_i(t)$ denote the neural state of node i, $u_i(t)$ the input signal for node i, $\boldsymbol{\psi}(t) := [\psi_1(t), \ldots, \psi_{|\mathcal{V}|}(t)]^\top$, and $\mathbf{u}(t) := [u_1(t), \ldots, u_{|\mathcal{V}|}(t)]^\top$. The DCM neural network can then be described by

$$\dot{\boldsymbol{\psi}}(t) = \delta \mathbf{A}\boldsymbol{\psi}(t) + \mathbf{u}(t) \tag{2.24}$$

where $\mathbf{A}$ stands for the network matrix corresponding to the ground truth, and δ is a scalar that adjusts the memory of the transition operator $\mathbf{A}$ (let $\delta = 20$ hereafter). Each nodal time series $\{\psi_i(t)\}$ obtained as a solution to (2.24) is subsequently fed into the nonlinear so-termed balloon model for vascular dynamics [39]. At the output, each simulated time

Toolboxes used in this thesis include [3, 36–38].

Algorithm 1 Multi-kernel learning for partial correlations

Require: $\boldsymbol{\theta}_0$, λ, Λ, η, $\{\kappa_p\}_{p=1}^P$.

for $(i,j) \in \mathcal{V} \times \mathcal{V}$, $(i < j)$, **do**
  for $l = 1, 2$ **do**
    $\mathbf{K}_{0\backslash ij} := \sum_{p=1}^P [\boldsymbol{\theta}_0]_p \mathbf{K}_{p\backslash ij}$.
    $\hat{\boldsymbol{\beta}}_i := (\mathbf{K}_{0\backslash ij} + \lambda \mathbf{I}_T)^{-1} \mathbf{x}_i$.
    while $\|\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i\|_2 \geq \epsilon_{acc.}$ **do**
      $\boldsymbol{\beta}_i := \hat{\boldsymbol{\beta}}_i$.
      $\mathbf{v} := [\boldsymbol{\beta}_i^\top \mathbf{K}_{1\backslash ij} \boldsymbol{\beta}_i, \ldots, \boldsymbol{\beta}_i^\top \mathbf{K}_{P\backslash ij} \boldsymbol{\beta}_i]^\top$.
      $\boldsymbol{\theta} := \boldsymbol{\theta}_0 + \Lambda \mathbf{v} / \|\mathbf{v}\|_2$.
      $\mathbf{K}_{\backslash ij}^{(i)} := \sum_{p=1}^P \theta_p \mathbf{K}_{p\backslash ij}$.
      $\hat{\boldsymbol{\beta}}_i := \eta \boldsymbol{\beta}_i + (1-\eta)(\mathbf{K}_{\backslash ij}^{(i)} + \lambda \mathbf{I}_T)^{-1} \mathbf{x}_i$.
    end while
    $i \leftrightarrow j$.
  end for
  $\hat{\mathbf{x}}_i := \mathbf{K}_{\backslash ij}^{(i)} \hat{\boldsymbol{\beta}}_i$.
  $\hat{\mathbf{x}}_j := \mathbf{K}_{\backslash ij}^{(j)} \hat{\boldsymbol{\beta}}_j$.
end for

series per node i is downsampled with period $TR = 3s$ to obtain the ith node data vector $\mathbf{x}_i$, comprising $T = 200$ samples.

In order to simulate RS-fMRI data, $u_i(t) = b_i(t) + n_i(t)$ is simulated as in [4], where $b_i(t)$ corresponds to a binary pulse train (20% average duty cycle) generated by a Markov chain, and $n_i(t)$ denotes zero-mean additive white Gaussian noise of variance 10^{-2}. Matrix $\mathbf{A}$ is chosen to be upper triangular, of dimensions 30×30, with fixed diagonal entries $\mathbf{A}_{ii} = -1$, and 100 randomly placed non-zero entries, each drawn uniformly at random from the interval $[0.25, 0.6]$. The signal decay rate and flow-dependent elimination rate parameters were set as the DCM priors in [13], while the remaining balloon model parameters were specified using a fit of experimental BOLD signals obtained under a field of 3 Tesla to the model, as detailed in [40].

The dictionary of kernels for the rest of this section consists of a single linear kernel and 19 Gaussian kernels (cf. (2.5)) with variances $\{\sigma_p^2\}_{p=1}^{19}$ belonging to the interval $[10^{-6}, 1]$. Finally, for each pair of nodes (i, j), the regularization parameters λ and Λ were selected separately using five-fold cross-validation from the grid $\{0.1, 1, 10, 100\} \times \{10, 50, 100\}$.

Testing KPC-based connectivity models

The proposed approach was evaluated with the DCM based synthetics using the empirical receiver operating characteristic (ROC) curves, which were obtained using $|\hat{\rho}_{ij|\mathcal{S}}|$ as a test statistic and gradually decreasing the threshold to vary the false alarm (FA) rate. The proposed KPC achieves a significant improvement in edge detection performance over linear PC, as confirmed by Fig. 1. Correctly identifying 70% of the ground truth edges in $\mathbf{A}$, results in 13 FAs when the proposed approach is used, as compared to 69 FAs for linear PC. As expected, upon fixing a maximum FDR level and applying the methods of Sec. 2.1.3 and in particular the procedure in [30] yields similar results (see Table 2.1).

In order to test the performance of the ε-insensitive variant proposed in Sec. 2.1.1, uniform noise was added on top of the DCM based synthetics described in Sec. 2.3.1, for a network consisting of $|\mathcal{V}| = 30$ nodes with 42 edges. The parameters of the algorithm were chosen as per Sec. 2.3.1, except for $\lambda = 10$ and $\Lambda = 100$. Fig. 2 plots the results

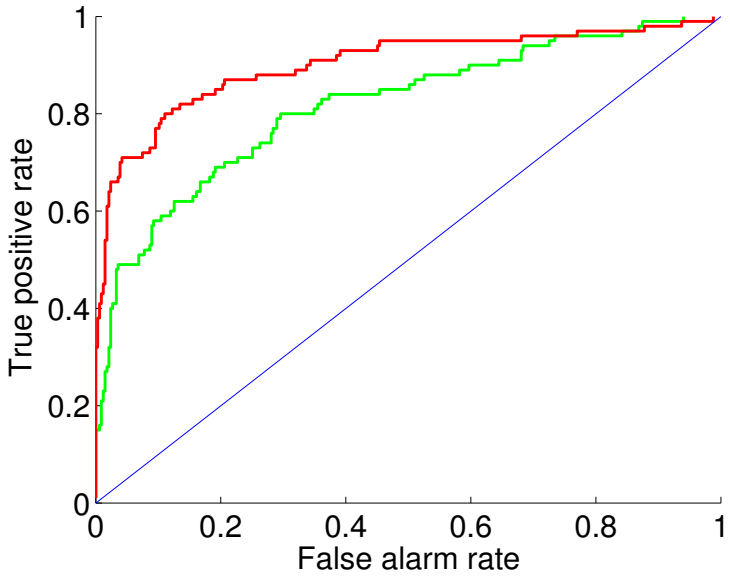

Figure 2.1: ROC curves obtained on DCM based synthetics. The red curve corresponds to the proposed kernel-based PC approach whereas the green one to linear PC.

	DCM	
	TPR(%)	FDR (%)
Proposed approach	65	10.96
Linear PC	49	25.76

Table 2.1: Performance of kernel-based and linear PC on DCM based synthetics obtained by thresholding the resulting PC matrices for a desired maximum FDR level of 0.15.

obtained for $\varepsilon = 0.02$ (one can readily employ cross-validation to choose ε). As expected, the ε-insensitive loss based KPC approach outperforms its counterpart based on the square loss. More importantly though, both methods outperform linear PC by a large margin in the uniform noise case as well.

Testing KPGC-based connectivity models

In order to assess performance of the proposed KPGC approach, synthetic data were generated using the forward DCM, sampled with $TR = 0.5$s. In particular, 7 random networks

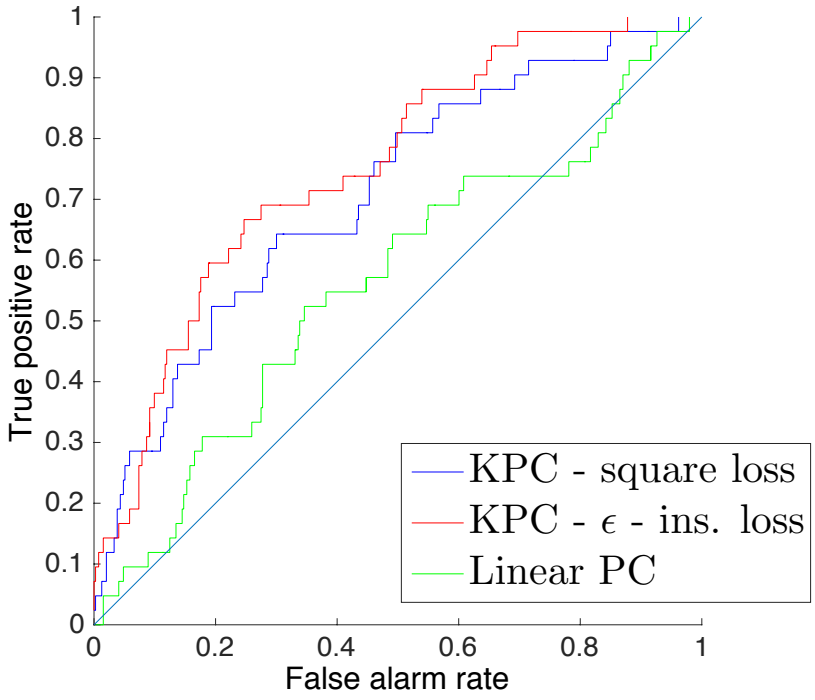

Figure 2.2: ROC curves obtained on DCM based synthetics, corrupted by uniform noise, using the proposed KPC, the ε-insensitive loss variant thereof and linear PC.

were generated as described in Sec. 2.3.1, each consisting of $|\mathcal{V}| = 10$ nodes. The average number of edges was 6.1, and each nodal vector was of length $T = 190$ samples. Model order selection for the novel KPGC method was performed using cross-validation, whereas for the linear and nonlinear PGC the Akaike Information Criterion was employed as described in [6]. Finally, 100 bootstrap realizations were used to estimate the empirical distribution of the PGC test statistics in (2.13), and correction for multiple testing was performed using the procedure in [31].

The figure of merit used was the ratio of the number of directed edges correctly identified over the number of possible directed edges dictated by the nonzero entries of the ground truth matrix $\mathbf{A}$. The proposed approach achieves an accuracy of 62%, outperforming both the linear and the nonlinear PGC [6], whose accuracy was close to chance (50%); see Table 2.2. Moreover, KPGC performed at least as good as the existing PGC methods in all individual simulations as well.

	Accuracy (%)
Proposed kernel-based PGC	62
Nonlinear PGC [6]	51
Linear PGC [6]	53

Table 2.2: Accuracy in identifying the presence of directed ground truth edges in DCM based synthetics by linear and nonlinear PGC, as well as by the proposed KPGC scheme.

2.3.2 Real data

RS-fMRI data obtained from 5 subjects were used in order to test performance of the proposed approaches on real data. The data obtained from 15 min scans were parcellated in 66 anatomically defined ROIs, and a time series of length $T = 445$ samples was associated with each ROI; see [41] for detailed description.

Testing KPC-based connectivity

For the subsequent test, a single kernel-based PC matrix with (i, j) entries $\{\hat{\rho}_{ij|\mathcal{S}}\}$ was obtained by averaging the corresponding matrices estimated for each subject. Further, by thresholding the aforementioned average matrix entries as described in Sec. 3.1 and using the procedure in [30], for a desired maximum FDR level $\alpha = 0.2$, a binary graph was obtained with $N = 202$ edges.

The networks estimated via the proposed KPC approach demonstrate widely observed [8, 42, 43] features of resting-state networks. First, the networks obtained exhibit small world characteristics. Let C_{KPC} and L_{KPC} denote the clustering coefficient and the average shortest path length of the estimated graph, respectively, while $C_{\text{E-R}}$ and $L_{\text{E-R}}$ denote the same quantities for a random Erdös-Renyi graph with the same number of nodes and edges; see also [3]. Our obtained graph will be deemed a small-world network if $L_{\text{KPC}} \simeq L_{\text{E-R}}$ and $C_{\text{KPC}} \gg C_{\text{E-R}}$ [44]. This was the case here, with $L_{\text{KPC}}/L_{\text{E-R}} = 1.2$ and $C_{\text{KPC}}/C_{\text{E-R}} = 3.1$. Recalling that our parcellation yielded 66 ROIs, inter-hemispheric links connecting 32 out of 33 regions with their homologous ones, were found. In addition, anatomically adjacent regions were found to be highly connected. Finally, the inferior parietal and the superior frontal cortex were the two nodes with the highest degree and shortest characteristic path length, whereas the precuneus was also found to be a high-degree node. The aforementioned ROIs are identified as hubs in RS networks [42].

Although the tests so far establish the plausibility of the estimated networks, what is worth stressing is that linear and nonlinear models give rise to distinct values of PC coefficients[3], as corroborated by Fig. 3, depicting $|\hat{\rho}_{ij|\mathcal{S}} - \hat{\rho}^{(l)}_{ij|\mathcal{S}}|$, where $\hat{\rho}^{(l)}_{ij|\mathcal{S}}$ denotes the (averaged across subjects) linear PC coefficient of $\mathbf{x}_i$ with $\mathbf{x}_j$.

It was also observed that the novel KPC method is able to better capture the underlying brain anatomy. In particular, we considered the task of inferring structural from functional connectivity [41], which is relevant since the latter is constrained by the former. Structural

[3]The graphical Lasso [45] was used to estimate the inverse covariance matrix (and thus the linear PC coefficients), since the sample covariance matrices for all subjects were found to be singular. The optimal value for the regularization parameter was chosen using cross-validation.

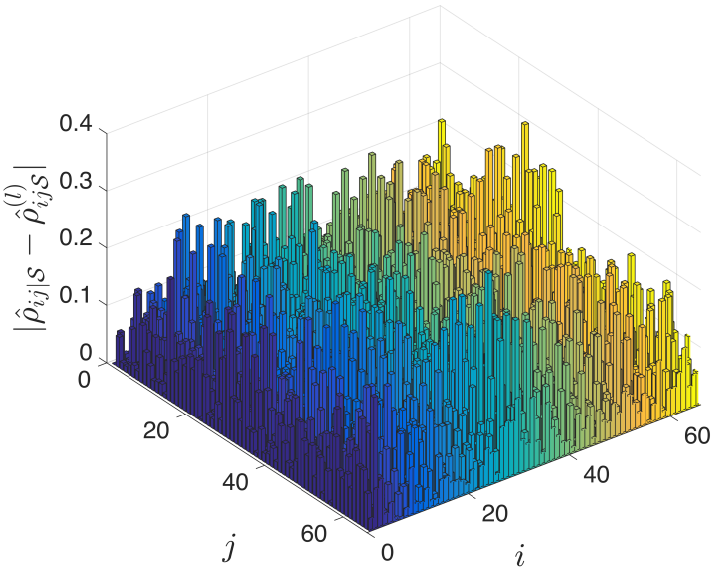

Figure 2.3: 3D bar graph of $|\hat{\rho}_{ij|\mathcal{S}} - \hat{\rho}^{(l)}_{ij|\mathcal{S}}|$ obtained on the real data described in Sec. 2.3.2.

connectivity for the 5 subjects was estimated using diffusion spectrum imaging tractography; see [41] for methods and details. As "ground-truth" for our functional connectivity (FC) matrix with entries $\{\hat{\rho}_{ij|\mathcal{S}}\}$, the 66×66 so-termed structural connectivity (SC) matrix was formed with (i, j) entries denoting the fiber densities between ROIs i, j [41]. The improvement achieved in detecting SC entries with KPC over linear PC is evident on the ROCs of Fig. 5[4].

Testing KPGC-based connectivity

The KPGC measure introduced in this thesis was used to assess the directionality of edges discovered by the KPC approach applied to data from a single subject[5]. When both potential directed edges for a particular undirected edge (i, j) were estimated to be absent, a

[4] A single functional connectivity (FC) matrix for each method was obtained by averaging across subjects the absolute values of the corresponding FC matrices.

[5] In particular, the first 5 mins of the scan were examined, and 50 bootstrap realizations were used to estimate the distribution of the PGC test statistic.

directed edge was declared according to $F_{ij} \underset{i \to j}{\overset{j \to i}{\gtrless}} F_{ji}$ where $j \to i$ denotes a directed edge from j to i, as per [7]. The results, for the right hemisphere, are summarized in Fig. 6. Given however the performance of KPGC on the DCM based synthetics, as well as the challenges associated with estimating Granger causality from the BOLD signal, the results should be treated with caution.

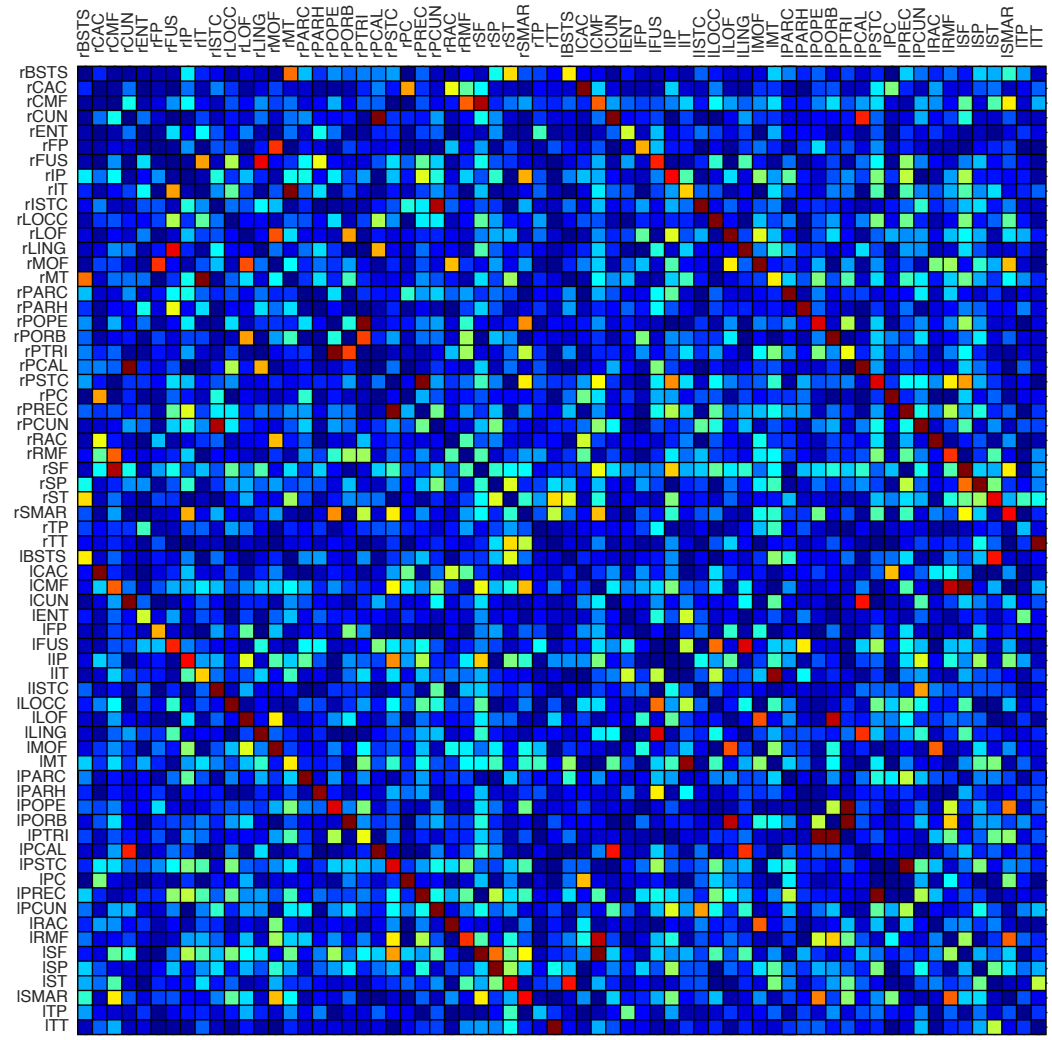

Figure 2.4: Representation of the functional network obtained using the kernel-based PCs on real RS-fMRI data (see Sec. 2.3.2). The absolute value of the KPC coefficient between a pair of regions is indicated by the color of the square in the corresponding position. High values are depicted with red, whereas lower ones are color-coded blue. Full names for the abbreviated regions of interest can be found in Appendix A.

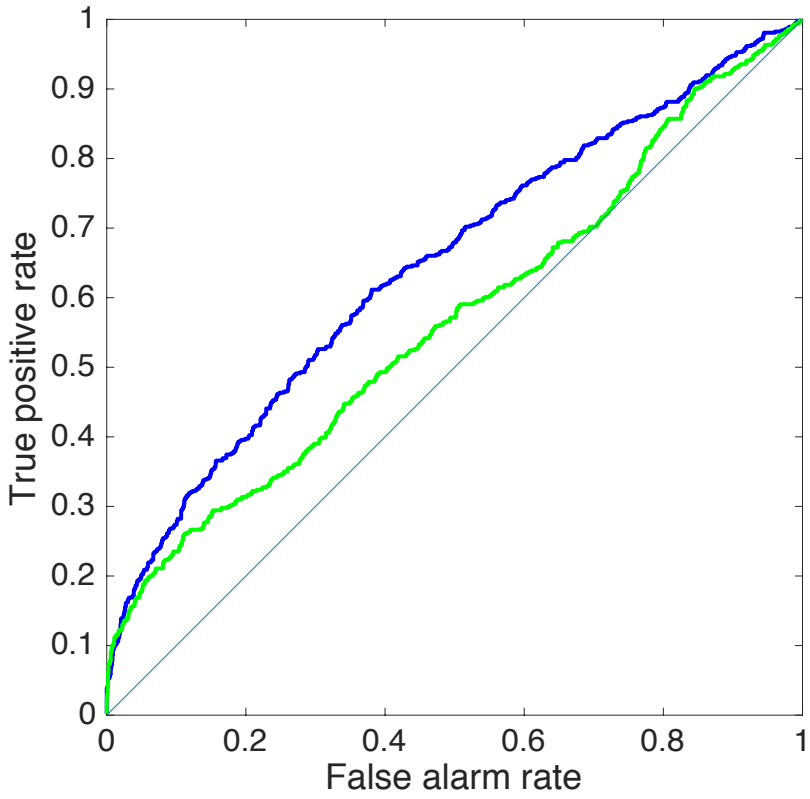

Figure 2.5: ROC curves obtained on inferring structural from functional connectivity, the latter being estimated using the proposed KPCs (blue curve) and linear PCs (green curve).

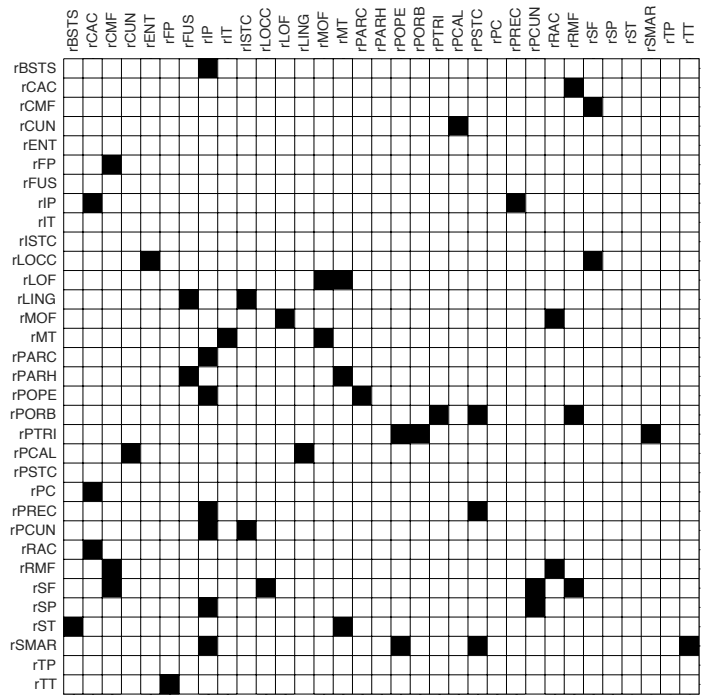

Figure 2.6: Edge directionality estimation on real RS fMRI data, using the novel KPGC approach. The presence of a directed edge $j \to i$ is visualized as a black square in position (i, j).

Chapter 3

Conclusion

3.1 Summary

Motivated by the limitations of linear brain connectivity models, and the presumed nonlinear nature of the BOLD signal, this thesis introduced a novel kernel-based nonlinear connectivity model to infer topology-revealing partial correlations (PCs), and partial Granger causality (PGC) measures. A data-driven multi-kernel approach was adapted to decipher the nonlinearity that optimally fits the data. Tests on DCM-based synthetic fMRI data demonstrated that the performance achieved by the kernel-based PC approach is superior to that of linear PC. Moreover, the proposed KPGC scheme was also seen to outperform both linear and nonlinear variants of PGC available, on such synthetics. Tests on real data reveal differences in the values of functional connectivity measures obtained using linear and nonlinear models, while at the same time they demonstrate the plausibility of networks estimated using the kernel-based PC approach. Finally, KPC based networks are found to be more reflective of the underlying brain anatomy as compared to linear PC based networks, which further speaks for the potential of the proposed approach.

3.2 Future directions

First of all, extensions of the proposed approach to the setting of dynamic functional connectivity, motivated by the observed performance thereof in low sample regimes, can be considered. An other interesting direction would be building on the proposed approach in order to develop (nonlinear) functional connectivity models that account for constraints imposed by structural connectivity. Finally, on an other front, wedding the merits of KPC in undirected topology identification with the ability of KPGC to estimate edge directionality would also be of interest.

Bibliography

[1] S. A. Huettel, A. W. Song, and G. McCarthy, *Functional Magnetic Resonance Imaging*. Sinauer Associates, 2004.

[2] O. Sporns, *Networks of the Brain*. MIT press, 2011.

[3] M. Rubinov and O. Sporns, "Complex network measures of brain connectivity: Uses and interpretations," *NeuroImage*, vol. 52, no. 3, pp. 1059–1069, 2010.

[4] S. M. Smith, K. L. Miller, G. Salimi-Khorshidi, M. Webster, C. F. Beckmann, T. E. Nichols, J. D. Ramsey, and M. W. Woolrich, "Network modelling methods for fMRI," *NeuroImage*, vol. 54, no. 2, pp. 875–891, 2011.

[5] G. Marrelec, A. Krainik, H. Duffau, M. Pélégrini-Issac, S. Lehéricy, J. Doyon, and H. Benali, "Partial correlation for functional brain interactivity investigation in functional MRI," *NeuroImage*, vol. 32, no. 1, pp. 228–237, 2006.

[6] S. Guo, A. K. Seth, K. M. Kendrick, C. Zhou, and J. Feng, "Partial Granger causality—Eliminating exogenous inputs and latent variables," *Journal of Neuroscience Methods*, vol. 172, no. 1, pp. 79–93, 2008.

[7] A. Roebroeck, E. Formisano, and R. Goebel, "Mapping directed influence over the brain using Granger causality and fMRI," *NeuroImage*, vol. 25, no. 1, pp. 230–242, 2005.

[8] S. Ryali, T. Chen, K. Supekar, and V. Menon, "Estimation of functional connectivity in fMRI data using stability selection-based sparse partial correlation with elastic net penalty," *NeuroImage*, vol. 59, no. 4, pp. 3852 – 3861, 2012.

[9] E. D. Kolaczyk, *Statistical Analysis of Network Data: Methods and Models*. Springer, 2009.

[10] N. K. Logothetis, J. Pauls, M. Augath, T. Trinath, and A. Oeltermann, "Neurophysiological investigation of the basis of the fMRI signal," *Nature*, vol. 412, no. 6843, Jul. 2001.

[11] W. Liao, D. Marinazzo, Z. Pan, Q. Gong, and H. Chen, "Kernel Granger causality mapping effective connectivity on fMRI data," *IEEE Trans. Medical Imaging*, vol. 28, no. 11, pp. 1825–1835, Nov. 2009.

[12] D. Marinazzo, M. Pellicoro, and S. Stramaglia, "Kernel method for nonlinear Granger causality," *Physical Review Letters*, vol. 100, no. 14, pp. 144–103, 2008.

[13] K. J. Friston, L. Harrison, and W. Penny, "Dynamic causal modelling," *NeuroImage*, vol. 19, no. 4, pp. 1273–1302, 2003.

[14] D. R. Hardoon, J. Mourao-Miranda, M. Brammer, and J. Shawe-Taylor, "Unsupervised analysis of fMRI data using kernel canonical correlation," *NeuroImage*, vol. 37, no. 4, pp. 1250–1259, 2007.

[15] E. Castro, V. Gómez-Verdejo, M. Martínez-Ramón, K. A. Kiehl, and V. D. Calhoun, "A multiple kernel learning approach to perform classification of groups from complex-valued fMRI data analysis: Application to schizophrenia," *NeuroImage*, vol. 87, pp. 1–17, 2014.

[16] B. Jie, D. Zhang, W. Gao, Q. Wang, C.-Y. Wee, and D. Shen, "Integration of network topological and connectivity properties for neuroimaging classification," *IEEE Trans. Biomed. Eng.*, vol. 61, no. 2, pp. 576–589, 2014.

[17] T. Zhang, D. Zhu, X. Jiang, B. Ge, X. Hu, J. Han, L. Guo, and T. Liu, "Predicting cortical ROIs via joint modeling of anatomical and connectional profiles," *Medical Image Analysis*, vol. 17, no. 6, pp. 601–615, Aug. 2013.

[18] G. V. Karanikolas, G. B. Giannakis, K. Slavakis, and R. M. Leahy, "Multi-kernel based nonlinear models for connectivity identification of brain networks," in *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing*, Shanghai, China, Mar. 2016, pp. 6315–6319.

[19] B. Schölkopf and J. A. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.* MIT press, 2002.

[20] B. Schölkopf, R. Herbrich, and A. J. Smola, "A generalized representer theorem," in *Computational Learning Theory: 14th Annual Conf. on Computational Learning The-*

ory, COLT and 5th European Conf. on Computational Learning Theory, EuroCOLT. Amsterdam, The Netherlands: Springer Berlin Heidelberg, Jul. 2001, pp. 416–426.

[21] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, no. 3, pp. 199–222, 2004.

[22] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1," 2014.

[23] ——, "Graph implementations for nonsmooth convex programs," in *Recent Advances in Learning and Control*, ser. Lecture Notes in Control and Information Sciences, V. Blondel, S. Boyd, and H. Kimura, Eds. Springer-Verlag Limited, 2008, pp. 95–110.

[24] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[25] A. B. Barrett, L. Barnett, and A. K. Seth, "Multivariate granger causality and generalized variance," *Phys. Rev. E*, vol. 81, p. 041907, 2010.

[26] C. W. J. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica*, vol. 37, no. 3, pp. 424–438, 1969.

[27] K. Friston, "Causal modelling and brain connectivity in functional magnetic resonance imaging," *PLoS Biology*, vol. 7, no. 2, p. e1000033, 2009.

[28] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*, 3rd ed. Wiley, 2003.

[29] S. Kay, *Fundamentals of Statistical Signal Processing, Volume II: Detection Theory*. Prentice Hall, 1998.

[30] Y. Benjamini and D. Yekutieli, "The control of the false discovery rate in multiple testing under dependency," *The Annals of Statistics*, vol. 29, no. 4, pp. 1165–1188, 2001.

[31] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.

[32] D. N. Politis and J. P. Romano, "The stationary bootstrap," *Journal of the American Statistical Association*, vol. 89, no. 428, pp. 1303–1313, 1992.

[33] D. N. Politis and H. White, "Automatic block-length selection for the dependent bootstrap," *Econometric Reviews*, vol. 23, no. 1, pp. 53–70, 2004.

[34] M. Gönen and E. Alpaydın, "Multiple kernel learning algorithms," *Journal of Machine Learning Research*, vol. 12, pp. 2211–2268, 2011.

[35] C. Cortes, M. Mohri, and A. Rostamizadeh, "L2 regularization for learning kernels," in *Proc. Conf. on Uncertainty in Artificial Intelligence*, ser. UAI '09, Arlington, VA, USA, 2009, pp. 109–116.

[36] T. Hayfield and J. S. Racine, "Nonparametric econometrics: The np package," *Journal of Statistical Software*, vol. 27, no. 5, 2008.

[37] A. Canty and B. D. Ripley, *boot: Bootstrap R (S-Plus) Functions*, 2016, r package version 1.3-18.

[38] A. C. Davison and D. V. Hinkley, *Bootstrap Methods and their Applications*. Cambridge University Press, 1997.

[39] R. B. Buxton, E. C. Wong, and L. R. Frank, "Dynamics of blood flow and oxygenation changes during brain activation: The balloon model," *Magnetic Resonance in Medicine*, vol. 39, no. 6, pp. 855–864, 1998.

[40] T. Mildner, D. G. Norris, C. Schwarzbauer, and C. J. Wiggins, "A qualitative test of the balloon model for BOLD-based MR signal changes at 3T," *Magnetic Resonance in Medicine*, vol. 46, no. 5, pp. 891–899, 2001.

[41] C. J. Honey, O. Sporns, L. Cammoun, X. Gigandet, J. P. Thiran, R. Meuli, and P. Hagmann, "Predicting human resting-state functional connectivity from structural connectivity," *Proc. of the Ntl. Academy of Sciences*, vol. 106, no. 6, pp. 2035–2040, 2009.

[42] S. Achard, R. Salvador, B. Whitcher, J. Suckling, and E. Bullmore, "A resilient, low-frequency, small-world human brain functional network with highly connected association cortical hubs," *The Journal of Neuroscience*, vol. 26, no. 1, pp. 63–72, 2006.

[43] R. Salvador, J. Suckling, M. R. Coleman, J. D. Pickard, D. Menon, and E. Bullmore, "Neurophysiological architecture of functional magnetic resonance images of human brain," *Cerebral Cortex*, vol. 15, no. 9, pp. 1332–1342, 2005.

[44] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[45] J. Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical Lasso," *Biostatistics*, vol. 9, no. 3, pp. 432–441, 2008.

[46] J. Cabral, E. Hugues, O. Sporns, and G. Deco, "Role of local network oscillations in resting-state functional connectivity," *NeuroImage*, vol. 57, no. 1, pp. 130–139, 2011.

Appendices

Appendix A

Full names of abbreviated ROIs

BSTS	Bank of the superior temporal sulcus
CAC	Caudal anterior cingulate cortex
CMF	Caudal middle frontal cortex
CUN	Cuneus
ENT	Entorhinal cortex
FP	Frontal pole
FUS	Fusiform gyrus
IP	Inferior parietal cortex
ISTC	Isthmus of the cingulate cortex
IT	Inferior temporal cortex
LING	Lingual gyrus
LOCC	Lateral occipital cortex
LOF	Lateral orbitofrontal cortex
MOF	Medial orbitofrontal cortex
MT	Middle temporal cortex
PARC	Paracentral lobule
PARH	Parahippocampal cortex
PC	Posterior cingulate cortex
PCAL	Pericalcarine cortex
PCUN	Precuneus
POPE	Pars opercularis
PORB	Pars orbitalis
PREC	Precentral gyrus

PSTC	Postcentral gyrus
PTRI	Pars triangularis
RAC	Rostral anterior cingulate cortex
RMF	Rostral middle frontal cortex
SF	Superior frontal cortex
SMAR	Supramarginal gyrus
SP	Superior parietal cortex
ST	Superior temporal cortex
TP	Temporal pole
TT	Transverse temporal cortex

Table A.1: Full names of the abbreviated ROIs in the main text. The prefixes l, r indicate the hemisphere to which each region belongs. Adapted from [46].

Appendix B

Multi-kernel learning for KPGC

Algorithm 2 Multi-kernel learning algorithm, for partial Granger causality, for assessing $i \to j$

Require: $\boldsymbol{\theta}_0$, λ, Λ, η, $\{\kappa_p\}_{p=1}^{P}$.

for $\mathcal{C} = \{i|V\backslash j, i|V\}$ **do**
  $\mathbf{K}_{0\mathcal{C}} := \sum_{p=1}^{P} [\boldsymbol{\theta}_0]_p \mathbf{K}_{p\mathcal{C}}$.
  $\hat{\boldsymbol{\beta}}_{\mathcal{C}} := (\mathbf{K}_{0\mathcal{C}} + \lambda \mathbf{I}_T)^{-1} \mathbf{x}_i^{(d)}$.
  while $\|\hat{\boldsymbol{\beta}}_{\mathcal{C}} - \boldsymbol{\beta}_i\|_2 \geq \epsilon_{acc.}$ **do**
    $\boldsymbol{\beta}_i := \hat{\boldsymbol{\beta}}_{\mathcal{C}}$.
    $\mathbf{v} := [\boldsymbol{\beta}_i^\top \mathbf{K}_{1\mathcal{C}} \boldsymbol{\beta}_i, \ldots, \boldsymbol{\beta}_i^\top \mathbf{K}_{P\mathcal{C}} \boldsymbol{\beta}_i]^\top$.
    $\boldsymbol{\theta} := \boldsymbol{\theta}_0 + \Lambda \mathbf{v} / \|\mathbf{v}\|_2$.
    $\mathbf{K}_{\mathcal{C}} := \sum_{p=1}^{P} \theta_p \mathbf{K}_{p\mathcal{C}}$.
    $\hat{\boldsymbol{\beta}}_{\mathcal{C}} := \eta \boldsymbol{\beta}_i + (1-\eta)(\mathbf{K}_{\mathcal{C}} + \lambda \mathbf{I}_T)^{-1} \mathbf{x}_i^{(d)}$.
  end while
end for
$\epsilon_{i|V\backslash j} = \mathbf{x}_i^{(d)} - \mathbf{K}_{i|V\backslash j} \hat{\boldsymbol{\beta}}_{i|V\backslash j}$.
$\epsilon_{i|V} = \mathbf{x}_i^{(d)} - \mathbf{K}_{i|V} \hat{\boldsymbol{\beta}}_{i|V}$