EMPIRICAL STATIONARY CORRELATIONS FOR SEMI-SUPERVISED LEARNING ON GRAPHS

By

Ya Xu
Justin S. Dyer
Art B. Owen

Technical Report No. 2009-7
May 2009

Department of Statistics
STANFORD UNIVERSITY
Stanford, California 94305-4065
EMPIRICAL STATIONARY CORRELATIONS FOR SEMI-SUPERVISED LEARNING ON GRAPHS

By

Ya Xu
Justin S. Dyer
Art B. Owen
Department of Statistics
Stanford University

Technical Report No. 2009-7
May 2009

This research was supported in part by
National Science Foundation grant DMS 0604939.

Department of Statistics
STANFORD UNIVERSITY
Stanford, California 94305-4065

http://statistics.stanford.edu
Empirical stationary correlations for semi-supervised learning on graphs

Ya Xu
Department of Statistics
Stanford University

Justin S. Dyer
Department of Statistics
Stanford University

Art B. Owen
Department of Statistics
Stanford University

January 2009

Abstract
In semi-supervised learning on graphs, response variables observed at one node are used to estimate missing values at other nodes. The methods exploit correlations between nearby nodes in the graph. In this paper we prove that many such proposals are equivalent to kriging predictors based on a covariance matrix driven by the link structure of the graph. We then propose a data-driven estimator of the correlation structure that exploits patterns among the observed response values. By incorporating even a small fraction of observed covariation into the predictions we are able to obtain much improved prediction on two graph datasets.

1 Introduction
Data on graphs has long been with us, but the recent explosion of interest in social network data available on the Internet has brought this sort of data to prominence. A typical problem is to predict the value of a feature at one or more nodes in the graph. That feature is assumed to have been measured on some, but not all nodes of the graph. For example, we might want to predict which web pages are spam, after a human expert has labeled a subset of them as spam or not. Similarly, we might want to know on which FaceBook web pages an ad would get a click, although that ad has only been shown on a subset of pages.
The underlying assumption in these prediction problems is that there is some correlation, usually positive, between the values at vertices that are close to each other in the graph. By making predictions that are smooth with respect to a notion of distance in the graph, one is able to define a local average prediction.

This problem is often called semi-supervised learning, because some of the data have measured response values, while others have predictor only. We suppose that the response random variable at node \( i \) of the graph is \( Y_i \). The observed value \( y_i \) is available at some, but not all \( i \). For a survey of semi-supervised learning see Zhu [2005].

Our starting point is the graph based random walk strategy of Zhou et al. [2005a]. To describe their approach, let \( G \) be a directed graph with \( n \) vertices. The edges of \( G \) are represented by an adjacency matrix \( W \) with entries \( w_{ij} > 0 \) if there is an edge from \( i \) to \( j \), and \( w_{ij} = 0 \) otherwise. We impose \( w_{ii} = 0 \), so that if the graph contains loops, we don’t count them. Node \( i \) has out-degree \( w_{i+} = \sum_{j=1}^{n} w_{ij} \) and in-degree \( w_{+i} = \sum_{j=1}^{n} w_{ji} \). The volume of the graph is \( w_{++} = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \).

There is a natural random walk associated with \( w_{ij} \) in which the probability of transition from \( i \) to \( j \) is \( P_{ij} = \frac{w_{ij}}{w_{i+}} \). Very often this walk is irreducible and aperiodic. If not, it may be reasonable to modify \( W \), by for example adding a small probability of a transition uniformly distributed on all nodes. For example, such a modification is incorporated into the PageRank algorithm of Page et al. [1998], to yield an irreducible and aperiodic walk on web pages.

An irreducible and aperiodic walk has a unique stationary distribution, that we call \( \pi \), which places probability \( \pi_i \) on vertex \( i \). Zhou et al. [2005a] define the similarity between \( i \) and \( j \) to be \( s_{ij} = \pi_i P_{ij} + \pi_j P_{ji} \). Then they construct a variation functional for vectors \( Z \in \mathbb{R}^n \) defined on the nodes of \( G \):

\[
\Omega(Z) = \frac{1}{2} \sum_{i,j} s_{ij} \left( \frac{Z_i}{\sqrt{\pi_i}} - \frac{Z_j}{\sqrt{\pi_j}} \right)^2. \tag{1}
\]

This variation penalizes vectors \( Z \) that differ too much over similar nodes. It also contains a scaling by \( \sqrt{\pi_i} \). One intuitive reason for such a scaling is that a small number of nodes with a large \( \pi_i \) could reasonably have more extreme values of \( Z_i \) while the usually much greater number of nodes with small \( \pi_i \) should not ordinarily be allowed to have very large \( Z_i \), and hence should be regularized more strongly. Mathematically, the divisors \( \sqrt{\pi_i} \) originate in spectral clustering and graph partitioning algorithms.
The prediction $Z$ should have a small value of $\Omega(Z)$. But it should also remain close to the observed values. To this end, they make a vector $Y^*$ where $Y_i^* = y_i$ when $y_i$ is observed and $Y_i^* = \mu_i$ when $y_i$ is not observed, where $\mu_i$ is a reasonable guess for $Y_i$. Then the predictions are given by

$$\hat{Y} = \arg\min_{Z \in \mathbb{R}^n} \Omega(Z) + \lambda \|Z - Y^*\|^2,$$

where $\lambda > 0$ is a parameter governing the trade off between fit and smoothness.

The minimizer $\hat{Y}$ in (2) is a linear function of $Y^*$. In very many of the applications $y_i \in \{-1, 1\}$ is binary and $Y_i^* = 0$ is used to represents uncertainty about their value. Although linear models may seem less natural than logistic regressions, they are often used for discrete responses because they are a computationally attractive relaxation of the problem of minimizing a quantity over $Z \in \{-1, 1\}^n$.

The outline of this paper is as follows. Because the random walk smoother leads to a linear method, we might expect it to have a representation as a minimum mean squared error linear prediction under a Gaussian process model for $Z$. That is, it might be a form of kriging. Section 2 presents notation for kriging methods. Section 3 exhibits a sequence of kriging predictors that converge to the random walk semi-supervised learning prediction (2).

The kriging model which yields random walk smoothing has a covariance assumption driven by the geometry of the graph in which the $\sqrt{\pi_i}$ values play a very strong role. Section 4 explores some other graph based semi-supervised learning methods which have different covariance assumptions in their corresponding kriging models. In Section 5 we derive another kriging method incorporating the empirical variogram of the observed $Y_i$ values into an estimate of the covariance. That method uses a full rank covariance, which is therefore computationally expensive for large $n$. We also present a lower rank version more suitable to large scale problems.

Section 6 presents two numerical examples. In Section 6.1, $Y_i$ is a numerical measure of the research quality of 110 universities in the UK and $w_{ij}$ measures the number of links from university $i$ to $j$. In holdout comparisons our kriging method is more accurate than the random walk smoother, which ends up being quite similar to linear regression on $\sqrt{\pi_i}$ values without an intercept. Section 6.2 presents a binary example, the WebKB dataset, where the response is 1 for student web pages and −1 otherwise. Incorporating empirical correlations into the semi-supervised learning methods brings a large improvement in the area under the ROC curve.
Section 7 describes some simple adaptations of the approach presented here. Section 8 has our conclusions.

2 Kriging on a graph

Kriging is named for the mining engineer Krige, whose paper Krige [1951] introduced the method. For background on kriging see Stein [1999]. Here we present the method and introduce the notation we need later.

A plain model for kriging on a graph works as follows. The data \( Y \in \mathbb{R}^n \) are written

\[
Y = \nu \beta + S + \varepsilon. \quad (3)
\]

Here \( \nu \in \mathbb{R}^{n \times k} \) is a matrix of predictors and \( \beta \in \mathbb{R}^k \) is a vector of coefficients. The structured part of the signal is \( S \sim \mathcal{N}(0, \Sigma) \) and it is the correlations within \( \Sigma \) that capture how neighbors in the graph are likely to be similar. Finally \( \varepsilon \sim \mathcal{N}(0, \Gamma) \) is measurement noise independent of \( S \). The noise covariance \( \Gamma \) is diagonal.

We adapt this model by taking \( \nu \in \mathbb{R}^n \), so \( k = 1 \), and then making \( \beta \) random from \( \mathcal{N}(\mu, \delta^{-1}) \), independent of both \( \varepsilon \) and \( S \). We write the result as

\[
Y = Z + \varepsilon \quad (4)
\]

where \( Z \sim \mathcal{N}(\mu \nu, \Psi) \), for \( \Psi = \nu \nu' \delta^{-1} + \Sigma \), independently of \( \varepsilon \sim \mathcal{N}(0, \Gamma) \).

In this formulation, the values \( Y \) that we have observed are noisy measurements of some underlying quantity \( Z \) that we wish we had observed. We seek to recover \( Z \) from measurements \( Y \).

Some of the \( Y_i \) are observed and some are not. None of the \( Z_i \) are observed. We let \( Y^{(0)} \) denote the random variables that are observed, and \( y^{(0)} \) be the values we saw for them. The unobserved part of \( Y \) is denoted by \( Y^{(1)} \). The kriging predictor is

\[
\hat{Z} = \mathbb{E}(Z \mid Y^{(0)}). \quad (5)
\]

Without loss of generality, suppose that the vectors are ordered with observed random variables before unobserved ones. We partition \( \Psi \) as follows

\[
\Psi = \begin{pmatrix}
\Psi_{00} & \Psi_{01} \\
\Psi_{10} & \Psi_{11}
\end{pmatrix} = \begin{pmatrix}
\Psi_{00} & \Psi_{01} \\
\Psi_{10} & \Psi_{11}
\end{pmatrix}.
\]
so that, for example, $\Psi_{00} = \text{cov}(Z^{(0)}, Z^{(0)})$ and $\Psi_\bullet_0 = \text{cov}(Z, Z^{(0)})$. The matrices $\Sigma$ and $\Gamma$, and the vector $\nu$, are partitioned the same way.

The joint distribution of $Z$ and $Y^{(0)}$ is

$$
\begin{pmatrix}
Z \\
Y^{(0)}
\end{pmatrix} \sim \mathcal{N}
\left(
\begin{pmatrix}
\mu

\mu_0
\end{pmatrix},
\begin{pmatrix}
\Psi & \Psi_\bullet_0 \\
\Psi_0 & \Psi_{00} + \Gamma_{00}
\end{pmatrix}
\right).
$$

Now we can write the kriging predictor explicitly as

$$
\hat{Z} = \Psi_\bullet_0 (\Psi_{00} + \Gamma_{00})^{-1} (y^{(0)} - \mu_0) + \mu_0
$$

$$
= (\nu_0 / \delta + \Sigma_\bullet_0)(\nu_0 / \delta + \Sigma_{00} + \Gamma_{00})^{-1} (y^{(0)} - \mu_0) + \mu_0,
$$

where $\nu_0$ contains components of $\nu$ corresponding to the $Y_i$ which were observed. In the special case where the whole vector $Y = y$ is observed, the kriging predictor is

$$
\hat{Z} = (\nu / \delta + \Sigma)(\nu / \delta + \Sigma + \Gamma)^{-1} (y - \mu) + \mu.
$$

Letting $\delta \to 0$ leads to a model with an improper prior in which $Y$ has infinite prior variance in the direction given by $\nu$. One natural choice for $\nu$ is the vector $1_n$ of all 1s. Then the mean response over the whole graph is not penalized, just fluctuations within the graph. We will see other natural choices for $\nu$.

Our approach to kriging is based on estimating $\Sigma$, $\Gamma$ and $\mu$, and then predicting by (6). The kriging approach also gives expressions for the variance of the prediction errors:

$$
\text{var}(Z \mid Y^{(0)} = y^{(0)}) = \Psi - \Psi_\bullet_0 (\Psi_{00} + \Gamma_{00})^{-1} \Psi_\bullet_0.
$$

The predictions do not necessarily interpolate the known values. That is $\hat{Z}^{(0)}_i$ need not equal $Y_i^{(0)}$. Instead some smoothing takes place. The predictions can be forced closer to the data by making $\Gamma_{00}$ smaller. One reason not to interpolate, is that when the graph correlations are strong, it may be possible to detect erroneous labels as cases where $|\hat{Z}^{(0)}_i - Y_i^{(0)}|$ is large.

### 3 Random walk smoothing as kriging

Here we show that random walk regularization can be cast in terms of a sequence of kriging estimators.
The random walk regularizer predicts the responses $Y$ by

$$
\hat{Y} = \arg\min_Z \frac{1}{2} \sum_{i,j} s_{ij} \left( \frac{Z_i}{\sqrt{\pi_i}} - \frac{Z_j}{\sqrt{\pi_j}} \right)^2 + \lambda \|Z - Y^*\|^2,
$$

(9)

where $Y_i^* = Y_i$ for observed values and $Y_i^* = \mu_i$ for the unobserved values. Zhou et al. [2005a] take $\mu_i = 0$ for unobserved $y_i \in \{-1, 1\}$.

Introduce the matrix $\Pi = \text{diag}(\pi_1, \ldots, \pi_n)$, let $s_{i+} = \sum_{j=1}^n s_{ij}$ and let $\tilde{\Delta}$ be the matrix with elements

$$
\tilde{\Delta}_{ij} = \begin{cases} 
s_{i+} - s_{ii} & \text{if } i = j \\
-s_{ij} & \text{if } i \neq j.
\end{cases}
$$

The matrix $\tilde{\Delta}$ is the graph Laplacian of $\tilde{G}$, which is our original graph $G$ after we replace the weights $w_{ij}$ by the similarities $s_{ij}$. We will also need the graph Laplacian $\Delta$ of $G$ with

$$
\Delta_{ij} = \begin{cases} 
w_{i+} - w_{ii} & \text{if } i = j \\
-w_{ij} & \text{if } i \neq j.
\end{cases}
$$

It is clear that $\Delta$ and $\tilde{\Delta}$ are symmetric matrices with an eigenvalue of 0 corresponding to eigenvector $1_n$. Assuming that a graph such as $G$ or $\tilde{G}$ is connected, its Laplacian is positive semi-definite and has rank $n - 1$ [Luxburg, 2007]. For later use we write

$$
\tilde{\Delta} = U' \text{diag}(d_1, d_2, \ldots, d_{n-1}, 0) U = \sum_{i=1}^n d_i u_i u_i'
$$

(10)

where $U'U = I_n$, with $d_i > 0$ for $i < n$ and $d_n = 0$.

In matrix terms, the right hand side of equation (9) is

$$
Z'\Pi^{-1/2}\tilde{\Delta}\Pi^{-1/2}Z + \lambda(Z - Y^*)'(Z - Y^*)
= Z'(\Pi^{-1/2}\tilde{\Delta}\Pi^{-1/2} + \lambda I)Z - 2\lambda Z'Y^* + \lambda Y'^*Y^*.
$$

For $\lambda > 0$ this is a positive definite quadratic form in $Z$ and we find that

$$
\hat{Y} = \lambda(\Pi^{-1/2}\tilde{\Delta}\Pi^{-1/2} + \lambda I)^{-1}Y^*
= (I + \lambda^{-1}\Pi^{-1/2}\tilde{\Delta}\Pi^{-1/2})^{-1}Y^*.
$$

(11)

Now we are ready to present the existing random walk algorithm as a special form of kriging. To get the random walk predictor (11), we
1) make strategic choices for $\Gamma$, $\Sigma$, and $\nu$,
2) treat the missing parts of $Y$ as observed,
3) use the full data kriging estimator (7), and then
4) take the limit as $\delta \to 0$ from above.

In detail, the recipe is as follows:

**Theorem 1.** Let $Y = Z + \varepsilon \in \mathbb{R}^n$. Suppose that $Z = \nu \beta + S$ where $\nu \in \mathbb{R}^n$, $\beta \sim N(\mu, 1/\delta)$ and $S \sim N(0, \Sigma)$. Let $\varepsilon \sim N(0, \Gamma)$ and assume that $S$, $\beta$, and $\varepsilon$ are mutually independent. Suppose that $Y^{(0)}$ comprising the first $r > 1$ elements of $Y$ is observed. Let $Y^* \in \mathbb{R}^n$ with $Y^*_i = Y^{(0)}_i$ for $i = 1, \ldots, r$ and $Y^*_i = \mu \nu_i$ for $i = r + 1, \ldots, n$. Let $\hat{Z}_\delta^*$ be the kriging estimator (7) applied with $y = Y^*$. Assume that the Laplacian matrix $\tilde{\Delta}$ derived for the similarity weighted graph $\tilde{G}$ on which $Y$ is defined satisfies (10). We now choose

$$
\Gamma = \lambda^{-1} I, \\
\Sigma = \Pi^{1/2} \tilde{\Delta}^+ \Pi^{1/2}, \\
\nu = (\sqrt{\pi_1}, \ldots, \sqrt{\pi_n})',
$$

where $\tilde{\Delta}^+$ is the Moore-Penrose inverse of $\tilde{\Delta}$ and $\Pi = \text{diag}(\pi_1, \ldots, \pi_n)$. Then

$$
\lim_{\delta \to 0^+} \hat{Z}_\delta^* = (I + \lambda^{-1} \Pi^{-1/2} \tilde{\Delta}^+ \Pi^{-1/2})^{-1} Y^*,
$$

which is the random walk predictor given by (11).

**Proof:** First we notice that $\nu = \Pi^{1/2} 1_n$. The kriging estimator is $\hat{Z}_\delta^* = M_\delta (Y^* - \mu \nu) + \mu \nu$, where

$$
M_\delta = \left( \frac{\nu \nu'}{\delta} + \Sigma \right) \left( \frac{\nu \nu'}{\delta} + \Sigma + \Gamma \right)^{-1}
= \left( \frac{\nu \nu'}{\delta} + \Pi^{1/2} \tilde{\Delta}^+ \Pi^{1/2} \right) \left( \frac{\nu \nu'}{\delta} + \Pi^{1/2} \tilde{\Delta}^+ \Pi^{1/2} + \lambda^{-1} I \right)^{-1}
= \Pi^{1/2} \left( 1_n 1_n' \delta \tilde{\Delta}^+ \right) \Pi^{1/2} \left( \Pi^{1/2} \left( \frac{1_n 1_n'}{\delta} + \tilde{\Delta}^+ \right) \Pi^{1/2} + \lambda^{-1} I \right)^{-1}. 
$$

(12)

The three matrices on the left of (12) are invertible. Moving their inverses inside the matrix inverse there, we get

$$
M_\delta = \left( I + \lambda^{-1} \Pi^{-1/2} \left( \frac{1_n 1_n'}{\delta} + \tilde{\Delta}^+ \right) \Pi^{-1/2} \right)^{-1}.
$$
Using (10) we write
\[
\frac{1}{\delta} \mathbf{1}_n \mathbf{1}_n' + \tilde{\Delta}^+ = U' \text{diag} \left( \frac{1}{d_1}, \frac{1}{d_2}, \ldots, \frac{1}{d_{n-1}}, \frac{n}{\delta} \right) U,
\]
noting that the last column of \( U \) is \( \pm \frac{1}{\sqrt{n}} \mathbf{1}_n \), the constant eigenvector. Now
\[
M_\delta = \left( I + \lambda^{-1/2} \Pi^{-1/2} U' \text{diag} \left( d_1, d_2, \ldots, d_{n-1}, \frac{\delta}{n} \right) U \Pi^{-1/2} \right)^{-1}.
\]
Letting \( \delta \to 0^+ \)
\[
M_\delta \to M_0 = (I + \lambda^{-1/2} \Pi^{-1/2} \tilde{\Delta} \Pi^{-1/2})^{-1}.
\]
This limit exists because the matrix being inverted is positive definite.

The terms related to the mean \( \mu \) vanish because
\[
(M_0 \nu - \nu) = (I + \lambda^{-1/2} \Pi^{-1/2} \tilde{\Delta} \Pi^{-1/2})^{-1} \Pi^{1/2} \mathbf{1}_n = \Pi^{1/2} \mathbf{1}_n
\]
\[
= (I + \lambda^{-1/2} \Pi^{-1/2} \tilde{\Delta} \Pi^{-1/2})^{-1} (\lambda^{-1/2} \Pi^{-1/2} \tilde{\Delta} \mathbf{1}_n + \Pi^{1/2} \mathbf{1}_n) - \Pi^{1/2} \mathbf{1}_n
\]
\[
= (I + \lambda^{-1/2} \Pi^{-1/2} \tilde{\Delta} \Pi^{-1/2})^{-1} (\lambda^{-1/2} \Pi^{-1/2} \tilde{\Delta} \Pi^{-1/2} + I) \Pi^{1/2} \mathbf{1}_n - \Pi^{1/2} \mathbf{1}_n
\]
\[
= 0.
\]
The first equality follows because \( \tilde{\Delta} \mathbf{1}_n = 0 \). Therefore, in view of (11), \( \tilde{Z}_\delta^* \to \tilde{Y} \) as \( \delta \to 0^+ \). □

One thing that stands out from the kriging analysis is the vector \( \nu = \sqrt{\pi} \) interpreted component-wise. The equivalent prior on \( Y \) in the direction parallel to \( \nu \) is improper. Thus the method anticipates that \( Y \) could be a large multiple of \( \nu \). When \( Y = \beta \nu \) for some value \( \beta \neq 0 \) the similar nodes are the ones with comparable values of \( \sqrt{\pi} \). These are not necessarily close together in the graph.

The next thing that stands out is that the correlation strength between nodes is a fixed property of the graph. If some response variables have stronger local correlations, others weaker, and still others negative local correlations, that is not reflected in this choice of \( \Sigma \).

4 Other semi-supervised learning as kriging

There are several other graph based semi-supervised learning methods that can be expressed in a similar regularization framework. In this section, we build a similar connection between some of these other semi-supervised
learning methods and kriging. The details are quite similar to the proof of Theorem 1 for random walk smoothing, and so we omit the proofs. Most of these examples are taken from the survey paper Zhu [2005]. Some of these methods were originally introduced with general loss functions, but we only consider their squared error loss versions. This is because $\hat{Y}$ may not be linear in $Y^*$ under a general loss function and hence is no longer kriging.

In each case there is a quadratic variation $\Omega(Z)$ and a quadratic error norm on $Z - Y^*$ each of which should ideally be small subject to a tradeoff between them. We take $\Omega(Z) = Z'LZ$ for a smoothing matrix $L$ and measure the error between $Z$ and $Y^*$ by $(Z - Y^*)'\Lambda(Z - Y^*)$. The smoothing matrix $L$ is positive semidefinite and $\Lambda$ is a diagonal matrix with $\Lambda_{ii} \geq 0$, while the sum $L + \Lambda$ is invertible. The algorithm then picks the minimizer of

$$Q(Z) = Z'LZ + (Z - Y^*)'\Lambda(Z - Y^*). \quad (13)$$

It is easy to show that

$$\hat{Y} \equiv \arg \min_Z Q(Z) = (L + \Lambda)^{-1}\Lambda Y^*. \quad (14)$$

For random walk smoothing in Section 3, $\Lambda$ is $\lambda I$ and $L$ is $\Pi^{-1/2}\Delta\Pi^{-1/2}$.

Random walk smoothing is defined for directed graphs. A few of the methods discussed below are only defined for undirected graphs. To apply one of them to a given directed graph, the standard technique is to work with $W + W'$.

We build the connection to kriging for several semi-supervised learning methods below. Then, to allow easy comparison of the methods, we present a summary in Table 1 at the end of this section.

4.1 Example one: Belkin et al. [2004]

Belkin et al. [2004] consider undirected graphs and use the (symmetric) edge weights $w_{ij}$ as similarities $s_{ij}$. Their Tikhonov regularization algorithm uses a criterion proportional to

$$Q(Z) = Z'\Delta Z + \lambda_0\|Z^{(0)} - Y^{(0)}\|^2,$$

with an option to use the side constraint $\frac{1}{n} \sum_{i=1}^n Z_i = \frac{1}{r} \sum_{i=1}^r Y_i^{(0)}$. That constraint forces the mean prediction to equal the mean observation, and is necessary for the generalization bound they obtained. We do not use this condition, because the squared error norm on $Z^{(0)} - Y^{(0)}$ already forces $Z^{(0)}$ to be close to $Y^{(0)}$. 

9
Their method fits the quadratic criterion (13) after making the substi-
tutions $L = \Delta$ and $\Lambda = \text{diag}(\lambda_0 I_r, 0 I_{n-r})$. The solution $\hat{Y}$ is the kriging estimator (7) with the following choices:

$$
\Gamma = \text{diag}(\lambda_0^{-1} I_r, \lambda_1^{-1} I_{n-r}),
\Sigma = \Delta^+ , \text{ and }
\nu = 1_n ,
$$

and taking limit as $\delta \to 0$ and then $\lambda_1 \to 0$.

There are two key differences between this method and random walk smoothing. First neither $\Sigma$ nor $\nu$ involve $\sqrt{\pi}$ here. Second, this model uses a diffuse prior on the noise for the unobserved responses, while random walk smoothing uses the same variance for both observed and unobserved responses.

Belkin et al. [2004] also propose an interpolating algorithm that leaves all the known values unchanged in the prediction. That is $\hat{Y}_i^{(0)} = Y_i^{(0)}$ for $i = 1, \ldots, r$. The resulting prediction arises in the limit as $\lambda_0 \to \infty$ for the Tikhonov estimator, and hence the connection to kriging remains the same.

They consider the generalization that replaces $\Delta$ by $\Delta^p$ for a positive integer power $p$. They also consider a generalization in which there could be more than one measurement made on the response variable at some of the nodes. We don’t consider cases more general than 0 or 1 observed response values per node.

**4.2 Example two: Zhou et al. [2004]**

Zhou et al. [2004] present an undirected graph algorithm that is a predecessor to the random walk smoothing of Zhou et al. [2005a]. For an undirected graph $w_{ij} = w_{ji}$, and of course the in– and out–degrees of each node coincide. Let $D$ be the diagonal matrix containing the common degree values $D_{ii} = w_{i+} = w_{+i}$.

They minimize

$$
\frac{1}{2} \sum_{i,j} w_{ij} \left( \frac{Z_i}{\sqrt{D_{ii}}} - \frac{Z_j}{\sqrt{D_{jj}}} \right)^2 + \lambda \| Z - Y^* \|^2
$$

(15)

which is the random walk smoothing criterion (9) after replacing the similarity $s_{ij}$ by the weight $w_{ij}$ and the stationary probability $\pi_i$ by the degree $D_{ii}$. Recall that for an irreducible aperiodic random walk on an undirected graph with transitions $P_{ij} = w_{ij}/w_{i+}$, the stationary distribution
has $\pi_i = D_{ii}/w_{++}$. Also, the similarity values become proportional to $w_{ij}$: 

$$s_{ij} = (D_{ii}/w_{++})(w_{ij}/D_{ii}) + (D_{jj}/w_{++})(w_{ji}/D_{jj}) = 2w_{ij}/w_{++}.$$ 

As a result, the symmetrized version of (9) is equivalent to (15) after multiplying $\lambda$ by $1/2$.

The criterion (15) fits the standard form (13) with $L = D^{-1/2}\Delta D^{-1/2}$ and $\Lambda = \lambda I$, where $\Delta$ is the graph Laplacian of $G$ (not $\tilde{G}$).

Their estimate reduces to the kriging estimator (7) with the following choices:

$$\Gamma = \lambda^{-1}I,$$

$$\Sigma = D^{1/2}\Delta^{1/2}D^{1/2},$$

and,

$$\nu = (\sqrt{D_{11}}, \ldots, \sqrt{D_{nn}}'),$$

in the limit as $\delta \to 0$.

### 4.3 Example three: Zhou et al. [2005b]

Zhou et al. [2005b] proposes another random walk based strategy on directed graphs that utilizes the hub and authority web model introduced by Kleinberg [1999]. For Zhou et al. [2005b], any node with an outlink is a hub and any node with an inlink is an authority. A node can be both a hub and an authority. They use two random walks. Their hub walk transitions between hubs that link to a common authority and their authority walk transitions between authorities linked by a common hub.

The hubs define a walk on the authorities as follows. From authority $i$ we pick a linking hub $h$ with probability $w_{hi}/w_{++}$ and from there pick an authority $j$ with probability $w_{hj}/w_{h+}$. The resulting transition probability from $i$ to $j$ is

$$P_{ij}^{(A)} = \sum_h \frac{w_{hi}}{w_{++}} \cdot \frac{w_{hj}}{w_{h+}}$$

where the sum is over hubs $h$. Analogous hub transition probabilities are

$$P_{ij}^{(H)} = \sum_a \frac{w_{ia}}{w_{a+}} \cdot \frac{w_{ja}}{w_{++}}$$

summing over authorities $a$.

The stationary distributions of these two walks have closed forms

$$\pi_i^{(H)} = w_{i+}/w_{++}, \quad \text{and} \quad \pi_i^{(A)} = w_{++}/w_{++}.$$
These formulas give appropriate zeros for nodes $i$ that are not hubs or authorities respectively.

We use stationary distributions and Laplacians of these two walks. Let $\Pi_H = \text{diag}(\pi_1^{(H)}, \ldots, \pi_n^{(H)})$ and $\Pi_A = \text{diag}(\pi_1^{(A)}, \ldots, \pi_n^{(A)})$. Then let $\tilde{\Delta}_H$ be the Laplacian of the graph $\tilde{G}_H$, which is our original graph $G$ after replacing the weights $w_{ij}$ by the similarity $s_{ij}^{(H)} = \pi_i^{(H)} P_{ij}^{(H)} + \pi_j^{(H)} P_{ji}^{(H)}$. Similarly let $\tilde{\Delta}_A$ be the Laplacian of $\tilde{G}_A$ which has weights $s_{ij}^{(A)} = \pi_i^{(A)} P_{ij}^{(A)} + \pi_j^{(A)} P_{ji}^{(A)}$.

The hub and authority regularization of Zhou et al. [2005b] uses the quadratic criterion (13) with $\Lambda = \lambda I$ and smoothing matrix

$$L = \gamma \Pi_H^{-1/2} \tilde{\Delta}_H \Pi_H^{-1/2} + (1 - \gamma) \Pi_A^{-1/2} \tilde{\Delta}_A \Pi_A^{-1/2}$$

for some $\gamma \in [0, 1]$. The choice of $\gamma$ allows the user to weigh the relative importance of inlinks and outlinks.

Their hub and authority walk smoother matches the kriging estimator (7) with the following choices:

$$\Gamma = \lambda I,$n
$$\Sigma = (\gamma \Pi_H^{-1/2} \tilde{\Delta}_H \Pi_H^{-1/2} + (1 - \gamma) \Pi_A^{-1/2} \tilde{\Delta}_A \Pi_A^{-1/2})^{-1},$$
$$\nu = 0_n.$$

Ordinarily, $L$ is positive definite for $0 < \gamma < 1$. The two terms in (16) each have one eigenvector with eigenvalue 0, but those two eigenvectors are, in general, linearly independent. We can construct exceptions. For example if $G$ is the complete graph then the hub and authority walks coincide and $L$ reduces to the random walk case which has one zero eigenvalue. More generally if every node has $w_{ii} = w_{i+1}$ the same thing happens. Outside of such pathological examples, $L$ is positive definite.

4.4 Example four: Belkin et al. [2006]

The manifold regularization framework introduced by Belkin et al. [2006] considers undirected graphs with similarity $s_{ij} = w_{ij}$. They predict the responses $Y$ by

$$\hat{Y} = \arg\min_Z \|Z\|_K^2 + \gamma Z' \Delta Z + \lambda_0 \|Z^{(0)} - Y^{(0)}\|^2,$$

where $K$ is a Mercer kernel (see [Cristianini and Shawe-Taylor, 2000, Chapter 3]), $\Delta$ is the graph Laplacian and $\gamma > 0$. The term $\|Z\|_K^2$ controls the smoothness of the predictions in the ambient space, while $Z' \Delta Z$ controls
the smoothness with respect to the graph. We consider the special case
where $K$ is a linear kernel. Then $\|Z\|^2_K = Z'KZ$ for a positive semidefinite
matrix $K \in \mathbb{R}^{n \times n}$. Now manifold regularization uses the criterion (13) with
$L = K + \gamma \Delta$ and $\Lambda = \text{diag}(\lambda_0 I_r, 0I_{n-r})$.

We have two cases to consider. The matrix $\gamma \Delta$ has $n - 1$ positive eigen-
values and an eigenvalue of 0 for the eigenvector $1_n$. If $K 1_n = 0$ then $L = K + \gamma \Delta$ is singular but otherwise $L$ is positive definite.

When $L$ is positive definite the implied prior is not improper in any d i rection
so we take $\nu = 0_n$. In this case, the manifold regularization predictions
are from the kriging estimator (7) with the following choices:

$$\Gamma = \begin{pmatrix} \lambda_0^{-1} I_r & 0 \\ 0 & \lambda_1^{-1} I_{n-r} \end{pmatrix},$$
$$\Sigma = (K + \gamma \Delta)^{-1}, \quad \text{and}$$
$$\nu = 0_n,$$

in the limit $\lambda_1 \to 0$.

Now suppose that $K + \gamma \Delta$ fails to be invertible because $K$ has eigenvector $1_n$ with eigenvalue 0. In this case, we replace $(K + \gamma \Delta)^{-1}$ by the
responding Moore-Penrose inverse and use $\nu = 1_n$, taking the limit $\delta \to 0$.

Our condition that the Mercer kernel be linear is necessary. For a general
Mercer Kernel $K$, the prediction $\hat{Y}$ need not be linear in $Y^*$, and so for such
kernels, manifold regularization does not reduce to kriging.

4.5 Other examples: smoothing matrix derived from $\Delta$

A few papers (e.g. Kondor and Lafferty [2002]; Smola and Kondor [2003];
Zhu et al. [2003]) construct the smoothing matrix $L$ based on a spectral
transformation of the graph Laplacian $\Delta$. They take

$$L = \sum_{i=1}^{n} f(d_i) u_i u_i', $$

where $d_i$ and $u_i$ are eigenvalues and eigenvectors of $\Delta$ as in (10), and $f(\cdot)$ is
a non-negative increasing function, such as $f(x) = e^{\alpha^2 x/2}$. 

13
When \( f(d_n) > 0 \), the connection to kriging can be written as
\[
\Gamma = \begin{pmatrix}
\lambda_0^{-1} I_r & 0 \\
0 & \lambda_1^{-1} I_{n-r}
\end{pmatrix},
\]
\[
\Sigma = \sum_{i=1}^{n} f(d_i) u_i u_i', \quad \text{and}
\]
\[
\nu = 0_n.
\]
For \( f(d_n) = 0 \), \( \Gamma \) remains the same but now
\[
\Sigma = \sum_{i=1}^{n-1} f(d_i) u_i u_i', \quad \text{and}
\]
\[
\nu = 1_n,
\]
with \( \delta \to 0 \).

5 Empirical stationary correlations

The semi-supervised learning methods presented in Section 4 all assume the signal covariance matrix \( \Sigma \) is a given function of the graph connectivity. In this section we show how to adapt the covariance to the dependency pattern seen among the non-missing \( Y \) values. We consider the random walk and Tikhonov predictors.

5.1 Stationary correlations

We begin with the Tikhonov regularization method because it is simpler. That method can be expressed in terms of kriging based on
\[
Y \sim \mathcal{N}(\mu 1_n, \Delta^+ + 1_n 1_n'/\delta + \Lambda^{-1})
\]
as \( \delta \to 0 \). The signal covariance \( \Delta^+ + 1_n 1_n'/\delta \) is a property of the graph. We replace it by \( \sigma^2 R \) where \( R \in \mathbb{R}^{n \times n} \) is an empirically determined correlation function that is stationary with respect to graph similarity. We take \( R_{ij} = \rho(s_{ij}) \) for a smooth function \( \rho \) that we estimate from the data as described below.

The random walk smoother corresponds to the kriging model
\[
Y \sim \mathcal{N}(\mu \nu, V(\Delta^+ + 1_n 1_n'/\delta) V + \Lambda^{-1})
\]
Table 1: Summary of connections between some semi-supervised learning methods and kriging.

<table>
<thead>
<tr>
<th>Reference</th>
<th>$\Gamma$</th>
<th>$\Sigma$</th>
<th>$\nu$</th>
<th>Limits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zhou et al. [2005a]</td>
<td>$\lambda^{-1}I$</td>
<td>$\Pi^{1/2} \tilde{\Delta}^{+} \Pi^{1/2}$</td>
<td>$\Pi^{1/2} \mathbf{1}_n$</td>
<td>$\delta \to 0$</td>
</tr>
<tr>
<td>(Random walk)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Belkin et al. [2004]</td>
<td>$(\lambda_0^{-1} I_r \ 0)_{0 \times n}$</td>
<td>$\Delta^+$</td>
<td>$\mathbf{1}_n$</td>
<td>$\delta \to 0$</td>
</tr>
<tr>
<td>(Tikhonov)</td>
<td></td>
<td></td>
<td></td>
<td>$\lambda_1 \to 0$</td>
</tr>
<tr>
<td>Belkin et al. [2004]</td>
<td>$(\lambda_0^{-1} I_r \ 0)_{0 \times n}$</td>
<td>$\Delta^+$</td>
<td>$\mathbf{1}_n$</td>
<td>$\delta \to 0$</td>
</tr>
<tr>
<td>(Interpolated)</td>
<td></td>
<td></td>
<td></td>
<td>$\lambda_1 \to 0$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$\lambda_0 \to \infty$</td>
</tr>
<tr>
<td>Zhou et al. [2004]</td>
<td>$\lambda^{-1}I$</td>
<td>$D^{1/2} \Delta^+ D^{1/2}$</td>
<td>$D^{1/2} \mathbf{1}_n$</td>
<td>$\delta \to 0$</td>
</tr>
<tr>
<td>Zhou et al. [2005b]</td>
<td>$\lambda^{-1}I$</td>
<td>$(1-\gamma)\Pi^{1/2} \tilde{\Delta}^{+} \Pi^{1/2}$</td>
<td>$\mathbf{1}_n$</td>
<td>$-$</td>
</tr>
<tr>
<td>(Hub &amp; authority)</td>
<td></td>
<td>$+\gamma \Pi^{1/2} \tilde{\Delta}^{+} \Pi^{1/2}$</td>
<td>$\mathbf{1}_n$</td>
<td>$-$</td>
</tr>
<tr>
<td>Belkin et al. [2006]</td>
<td>$(\lambda_0^{-1} I_r \ 0)_{0 \times n}$</td>
<td>$(K + \gamma \Delta)^{-1}$</td>
<td>$\mathbf{0}_n$</td>
<td>$\lambda_1 \to 0$</td>
</tr>
<tr>
<td>(Manifold, $K \mathbf{1}_n \neq \mathbf{0}_n$)</td>
<td>$K^{-1}$</td>
<td></td>
<td></td>
<td>$-$</td>
</tr>
<tr>
<td>Belkin et al. [2006]</td>
<td>$(\lambda_0^{-1} I_r \ 0)_{0 \times n}$</td>
<td>$(K + \gamma \Delta)^{-1}$</td>
<td>$\mathbf{1}_n$</td>
<td>$\delta \to 0$</td>
</tr>
<tr>
<td>(Manifold, $K \mathbf{1}_n = \mathbf{0}_n$)</td>
<td></td>
<td></td>
<td></td>
<td>$\lambda_1 \to 0$</td>
</tr>
<tr>
<td>Spectral transform</td>
<td>$f(d_n) &gt; 0$</td>
<td>$\sum_{i=1}^n f(d_i)^{-1} u_i u'_i$</td>
<td>$\mathbf{0}_n$</td>
<td>$-$</td>
</tr>
<tr>
<td>Spectral transform</td>
<td>$f(d_n) = 0$</td>
<td>$\sum_{i=1}^{n-1} f(d_i)^{-1} u_i u'_i$</td>
<td>$\mathbf{1}_n$</td>
<td>$\delta \to 0$</td>
</tr>
</tbody>
</table>

where $V = \text{diag}(\nu)$ and $\nu_i = \sqrt{\pi_i}$, as $\delta \to 0$. There is more than one way to inject empirically measured correlations into this model. We choose to once again replace $\Delta^+ + \mathbf{1}_n \mathbf{1}'_n / \delta$ by $\sigma^2 R$, so that the model becomes

$$Y \sim \mathcal{N}(\mu \nu, \sigma^2 V RV + \Lambda^{-1}).$$

It is implicit in this model that observations $Y_i$ with a larger absolute mean $|\nu_i|$ also have a larger signal variance $\sigma^2 \nu_i^2$. This seems reasonable but of course some data will be better fit by other relationships between mean and variance.

One appealing feature of model (17) is that it reduces to our empirical stationary correlation version of Tikhonov smoothing by taking $\nu = \mathbf{1}_n$. Moreover, (17) has a simple interpretation that it effectively models the
scaled signals $Z_i/\nu_i$ to be a Gaussian process with constant mean $\mu$ and stationary covariance $\sigma^2 R$. As a result, model (17) has a stationary signal correlation, but not necessarily a stationary mean or covariance.

Similar to the regularization methods discussed so far, we assume $\nu$ represents some prior information and is given. We also simplify the noise variance by assuming $\Lambda = \lambda I$. Notice that assuming a different noise variance for the unobserved $Y$ is unnecessary because it does not affect the kriging predictor (6) at all.

5.2 Covariance estimation through the variogram

Here we adapt the variogram-based approach from geostatistics (see for example Cressie [1993]) to estimate the matrix $R$. For $\Lambda = \lambda I$ the variogram of the model (17) is

$$\Phi_{ij} \equiv \frac{1}{2} \mathbb{E}((Y_i - \mu \nu_i) - (Y_j - \mu \nu_j))^2$$

$$= \lambda^{-1} + \frac{1}{2} \sigma^2 (\nu_i^2 + \nu_j^2 - 2\nu_i \nu_j R_{ij}).$$

(18)

For $1 \leq i, j \leq r$ both $Y_i = y_i$ and $Y_j = y_j$ are observed and so we have the naive estimator

$$\hat{\Phi}_{ij} = \frac{1}{2}((y_i - \mu \nu_i) - (y_j - \mu \nu_j))^2.$$  

(19)

The naive variogram is our starting point. We translate it into a naive value $\hat{R}_{ij}$ by solving equation (18). This requires known values of $\nu_i$ and $\nu_j$ which we also assume are non-zero. See Section 7 for the $\nu = 0_n$ cases. We also need values for $\lambda$ and $\sigma$. We will use cross-validation to choose $\lambda$ and $\sigma$, and so they also can be considered known at the time we are solving for $\hat{R}_{ij}$.

Once we have the naive correlation estimates $\hat{R}_{ij}$ we use a spline smoother to fit the smooth function $\hat{R}_{ij} = \hat{\rho}(s_{ij})$. Smoothing serves two purposes. It yields correlation as a function of similarity $s_{ij}$, and it reduces sampling fluctuations. Next we use $\hat{\rho}$ to estimate the entire correlation matrix via $\hat{R}_{ij} = \hat{\rho}(s_{ij})$ for $i \neq j$ with of course $\hat{R}_{ii} = 1$. To complete our estimation of the signal variance we take $\hat{\Sigma} = \sigma^2 V \hat{R} V$, and then if necessary modify it to be positive semi-definite.

The step-by-step procedure to estimate the signal covariance is listed in Table 2. The estimated $\Sigma$ can then be used in (6) to make predictions (with $\delta \to 0$).
1. For every pair of observed nodes \(i, j = 1, \ldots, r\) and \(i \neq j\), estimate \(R_{ij}\) by solving (18) with \(\Phi_{ij}\) estimated using (19):

\[
\hat{R}_{ij} = \frac{\sigma^2(\nu_i^2 + \nu_j^2)/2 + \lambda^{-1} - \hat{\Phi}_{ij}}{\sigma^2\nu_i \nu_j},
\]

(20)

2. Smooth the pairs \(\{(\hat{R}_{ij}, s_{ij}) : i, j = 1, \ldots, r\}\) to obtain the estimated correlation function \(\hat{\rho}(\cdot)\).

3. Compute \(\tilde{R}_{ij} = \hat{\rho}(s_{ij})\) for \(i \neq j\) and \(\tilde{R}_{ii} = 1\).

4. Set \(\hat{\Sigma} = \sigma^2 V \tilde{R} V\).

5. Pick one of the following two methods to make \(\hat{\Sigma}\) positive semi-definite. Let \(\hat{\Sigma} = U' H U\) be the eigen–decomposition of \(\hat{\Sigma}\). Then

(a) use \(\hat{\Sigma}_+ = U' H_+ U\), where \(H_+ = \max(H, 0)\), or,

(b) use \(\hat{\Sigma}_+^{(k)} = U' H_+^{(k)} U\), where \(H_+^{(k)}\) consists of the first \(k\) diagonal elements of \(H_+\) and the rest are set to be zero.

Choice (a) gives the positive semi-definite matrix that is closest to \(\hat{\Sigma}\) in Frobenius norm. Choice (b) is used when computational cost is a concern or the true covariance \(\Sigma\) is believed to be low-rank.

| 5.3 Relation to geostatistics |

We use a nonparametric estimate \(\hat{\rho}(\cdot)\) to avoid forcing a parametric shape on the correlation function. The classical parametric curves used in the geostatistics literature for \(\mathbb{R}^d\) may not work well for the graph and the similarity measure is not a distance in Euclidean space. Indeed it is more of an inverse distance.

Both Hall et al. [1994] and Shapiro and Botha [1991] have discussed ways to fit a nonparametric variogram while ensuring a positive semi-definite covariance. Their techniques apply when the predictor space is \(\mathbb{R}^d\). The usual definition of the similarity measure on graph is far from being a metric in \(\mathbb{R}^d\). Our approach ensures that the estimate for \(\Sigma\) is positive semi-definite.
When there are \( n \) observations, Hall et al. [1994] find convergence rates for the smoother \( \hat{\rho} \) that are comparable to that using \( n^2 \) observations. The reason is that we get \( O(n^2) \) pairs \((Y_i, Y_j)\) in the empirical variogram. In our application there are only \( r(r-1)/2 \) observed pairs to use.

In the spatial smoothing problems where kriging originates, it is often necessary for the covariance to remain semi-definite at any finite list of points in \( \mathbb{R}^d \), including some that are not yet observed. Our setting does not require us to construct an extension of the covariance function to \( Y_i \) for nodes \( i \) that are not in the graph. Even in cross-validation, we know the positions in the graph for the points for which no \( Y \) values have been observed and so we can still compute \( s_{ij} \) for all data pairs. This aspect of the semi-supervised setting makes the problem much simpler than that faced in geostatistics. It does however mean that when the graph changes, the covariance model may have to be recomputed.

6 Examples

We now compare our empirical covariance approach with the random walk and the Tikhonov regularization methods. We use two extremely different real datasets. The first one has a continuous response on a dense, weighted graph, and the second one has a binary response on a sparse, unweighted graph. Because both graphs are directed, we construct an undirected graph for the Tikhonov approach using \( W + W' \) as the adjacency matrix. Our empirical-based method, together with its low rank variations, brings substantial improvements for both methods on both datasets.

The random walk and the Tikhonov methods use very different choices for \( \nu \) and \( s_{ij} \), as discussed before. The choice of \( \nu \) corresponds to a prior belief in the direction that \( Y \) varies along, and \( s_{ij} \) defines the closeness of nodes \( i \) and \( j \). We believe their values should be dataset-dependent, but that is not the emphasis of this paper. Instead we use the default values of \( \nu \) and \( s_{ij} \) for each method and then focus on how performance changes when we incorporate empirical stationary correlations. Specifically, \( \nu = \sqrt{\pi} \) and \( s_{ij} = \pi_i P_{ij} + \pi_j P_{ji} \) for the random walk method, while \( \nu = 1_n \) and \( s_{ij} = w_{ij} + w_{ji} \) for the Tikhonov method.

For random walk smoothing, we need to estimate the overall mean \( \mu \). It disappears from the covariance in the \( \delta \to 0 \) limit, but it remains necessary in the construction of \( Y^* \). For binary problems with \( Y_i \in \{-1, 1\} \) we take \( \mu = 0 \), as is done in the machine learning literature. For continuous responses
we use

\[ \hat{\mu} = \frac{1}{r} \sum_{i=1}^{r} \frac{y_i}{\nu_i} \]  

We also investigated estimating \( \mu \) by generalized least squares regression of \( Y^{(0)} \) on \( \nu^{(0)} \) taking account of estimated correlations among the first \( r \) response values. This made only a very small difference even on the small problems we are about to report, and so we see no reason to prefer it to the very simple estimate (21).

For the Tikhonov method, we do not even need to estimate \( \mu \) at all. It disappears from the covariance in the \( \delta \to 0 \) limit and it disappears from \( Y^* \) in the \( \lambda_1 \to 0 \) limit.

6.1 The UK University Web Link Dataset

The university dataset contains the number of web links between UK universities in 2002. Each university is associated with a research score (RAE), which measures the quality of the university’s research\(^1\). After removing four universities that have missing RAE scores, or that have no in-link or out-link, there are 107 universities.

The response variable, RAE score, is continuous and ranges from 0.4 to 6.5 with a mean of 3.0 and a variance of 3.5. The number of links from one university to another forms the (asymmetric) weighted adjacency matrix \( W \).

The distribution of the weights \( w_{ij} \) is heavily right tailed and approximately follows a power law. About 15% of the weights are zero, and 50% of them are less than 7, while the maximum is 2130.

We first use the entire dataset to illustrate the empirical variance estimation procedure as listed in Table 2. Our performance numbers are based on cross-validated versions. For simplicity, we suppose \( \nu = 1_n \) and \( s_{ij} = w_{ij} + w_{ji} \), while using the tuning parameter values \( \sigma^2 = 5 \) and \( \lambda^{-1} = 0.01 \) that give the smallest residual sum of squares in fitting \( Y \).

Figure 1 (left) plots the naive estimates \( \hat{R}_{ij} \), as computed in (20), against (log transformed) similarity \( s_{ij} \) values. The logarithm is used because the \( s_{ij} \) are skewed. The scatter plot is very noisy, but we can nonetheless extract a non-trivial \( \hat{\rho}(\cdot) \) with cubic spline smoothing, as shown by the red curve. Here and elsewhere, we used cubic splines with ten knots to get \( \hat{\rho}(\cdot) \). The same curve is also included on the right plot at a larger scale. Note that \( \hat{\rho}(\cdot) \)

\(^1\)The are at http://cybermetrics.wlv.ac.uk/database/stats/data/. We use the link counts at the directory level.
is not monotonically increasing in $s_{ij}$. The greatest correlations arise for very similar nodes, but the very least similar node pairs also have somewhat more correlation than do pairs with intermediate similarity.

The next and final step in Table 2 is to positive semi-definitiz e the co-variance matrix $\hat{\Sigma}$ directly resulting from $\hat{\rho}(-)$. The full rank version gives $\hat{\Sigma}_+$ that is closest to $\hat{\Sigma}$ in terms of the Frobenius norm, and therefore we would expect $\hat{\Sigma}_+/\sigma^2$ to closely follow $\hat{\rho}(-)$ when $\nu = \mathbf{1}_n$. This is exactly the case here as shown in Figure 1 (right). Similarly, a low rank estimation $\hat{\Sigma}^{(k)}_+$ also scatters around $\hat{\rho}(-)$ but more loosely and its plot is omitted here.

Now we turn to performance comparisons. For this, we hold out the RAE scores of some universities and measure each prediction method by mean squared error (MSE) on the held out scores. The size of the holdout set ranges from approximately 10% to 90% of the entire dataset, and 50 trials are done at each holdout level.

Our method has two tuning parameters $\lambda$ and $\sigma$ while the graph based methods have only one. The comparison is fair though because it is based on hold out sets. For each set of held-out data we used ten-fold cross-validation within the held-in data to pick $\lambda$ and $\sigma$ for empirical stationary correlation kriging. For the plain random walk and Tikhonov methods we use the best tuning parameter ($\lambda$), and so our comparisons are to somewhat
better versions of the random walk and Tikhonov method than one could actually get in practice.

We found for this data that the random walk smoother is very close to a simple method that just regresses the responses $Y_i$ on $\nu_i = \sqrt{\pi_i}$. It hardly seems to use the graph-based similarities at all. Therefore we define a baseline method using $\hat{\mu}_i$ to predict $Y_i$ and we measure performance relative to this baseline. For the Tikhonov smoother $\nu_i = 1$ and the baseline is simply the average of measured responses.

The results are shown in Figure 2. The random walk method performs quite well compared to the Tikhonov method, but neither of them outperform their corresponding baseline methods by much, even with the best tuning parameters. The black and red curves track each other closely over a wide range of data holdout sizes, with the red (graph-based) curve just slightly lower than the black (baseline) curve.

The results show that the random walk choices $\nu = \sqrt{\pi_i}$ and $s_{ij} = \pi_i P_{ij} + \pi_j P_{ji}$ are clearly better than the Tikhonov choices $\nu = 1_n$ and $s_{ij} = w_{ij} + w_{ji}$ for the UK university data. A major difference between the methods is that the Tikhonov method symmetrizes the graph. As such, it does not distinguish between links from University $i$ to $j$ and links in the

Figure 2: This figure compares mean square error for predictions of held out RAE scores. The horizontal axis shows the number of university RAE scores held out. The vertical axis shows mean square prediction errors for the semi-supervised graph based methods described in the text.
Improvement over baseline

<table>
<thead>
<tr>
<th></th>
<th>Random walk</th>
<th>Tikhonov</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline MSE</td>
<td>1.71</td>
<td>3.64</td>
</tr>
<tr>
<td>Random walk</td>
<td>3.8%</td>
<td>-</td>
</tr>
<tr>
<td>Tikhonov</td>
<td>-</td>
<td>3.2%</td>
</tr>
<tr>
<td>Empirical</td>
<td>25.0%</td>
<td>50.9%</td>
</tr>
<tr>
<td>Empirical R5</td>
<td>32.4%</td>
<td>53.9%</td>
</tr>
<tr>
<td>Empirical R1</td>
<td>19.1%</td>
<td>50.9%</td>
</tr>
</tbody>
</table>

Table 3: This table presents the held out mean square prediction errors that are shown in Figure 2, when 50 of 107 university scores are held out. The baseline methods are simple regressions through the origin on $\nu = \sqrt{\pi}$ for the random walk case and on $\nu = 1_n$ for the Tikhonov case. The random walk and Tikhonov semi-supervised learning methods bring modest improvements over their respective baseline methods. Finally, the kriging methods using an empirical stationary correlation as described in the text, bring large improvements in the mean squared error.

other direction. Even the baseline for the random walk method, which does regression on $\sqrt{\pi}$, makes use of the directionality because that directionality is reflected within $\pi$.

The green curves in Figure 2 show the error rates for empirical stationary correlation kriging. They generally bring large performance improvements, except in the Tikhonov case when 70 to 90 of 107 observations are held out. In the latter case only about 17 University scores are being used and while this is probably too few to estimate a good covariance, it does not do much harm either. All the methods do better when less data are held out. The methods with data driven correlations have slightly steeper curves.

We make a numerical summary of the curves from Figure 2 in Table 3. We compare performance for the setting where about half of the data are held out. For both prediction methods, kriging with empirical stationary correlations typically brings quite large improvements. Low rank variations of empirical stationary correlation kriging perform similarly to the full rank empirical method, except for the rank 1 case in the random walk setting. There we still see a large improvement but not as much as for the full rank or rank 5 cases.
6.2 The WebKB Dataset

The WebKB dataset\(^2\) contains webpages collected from computer science departments of various universities in January 1997. The pages were manually classified into seven categories: student, faculty, staff, department, course, project and other. The dataset we have is a subset, where the webpages belonging to the “other” class are removed. We will only use the data for Cornell University, which has 195 webpages and 301 links, after removing the three self loops. We further reduce the webpage labels to be “student” (1) and “non-student” (−1). There are 83 student pages in total. The adjacency matrix is unweighted, i.e., \(w_{ij}\) is 1 if there is a link from page \(i\) to \(j\) and 0 otherwise. Again, the links are directed and hence \(W\) is asymmetric, with 99.2% of the \(w_{ij}\) being zero.

The kriging models make continuous predictions of the binary response. We use the area under the ROC curve (AUC) to measure performance on the holdout sets. The AUC is equivalent to the probability that a positive label will get a higher prediction than a negative label. To estimate the correlation function in the empirical based method, we again use cubic splines with ten knots for the random walk \(s_{ij}\). However, for the Tikhonov \(s_{ij}\), which has only three possible values 0, 1 and 2 in an unweighted directed graph, we simply use the average at each \(s_{ij}\) without smoothing. The tuning parameters are picked in the same way as for the university dataset.

The results are plotted in Figure 3 and summarized in Table 4. As a baseline, we consider a model which sorts the web pages in random order. It would have an AUC of 0.5. For the webKB data, the Tikhonov method has better accuracy than the random walk method which has trouble getting an AUC below 0.5, despite taking account of the edge directionality. In both cases empirical stationary correlations bring large improvements. As before we see that larger amounts of missing data make for harder prediction problems.

7 Variations

In many applications, we may want to use more nuanced error variance measures, such as \(\Gamma = \text{diag}(\sigma_1^2, \ldots, \sigma_n^2)\) and this fits easily into the kriging framework. For example, web pages determined to be spam after a careful examination could be given a smaller \(\sigma_i^2\) than those given less scrutiny, and those not investigated at all can be given a still higher \(\sigma_i^2\).

Figure 3: This figure compares semi-supervised learning methods described in the text. The response was an indicator of whether a webpage was a student page or not. The horizontal axis shows the number of labels held out and the vertical axis is 1 minus the area under the ROC curve.

<table>
<thead>
<tr>
<th>Improvement over baseline</th>
<th>Random walk</th>
<th>Tikhonov</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (1−AUC)</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>Random walk</td>
<td>−5.4%</td>
<td>-</td>
</tr>
<tr>
<td>Tikhonov</td>
<td>-</td>
<td>8.5%</td>
</tr>
<tr>
<td>Empirical</td>
<td>43.0%</td>
<td>37.5%</td>
</tr>
<tr>
<td>Empirical R5</td>
<td>40.0%</td>
<td>31.9%</td>
</tr>
<tr>
<td>Empirical R1</td>
<td>29.0%</td>
<td>16.3%</td>
</tr>
</tbody>
</table>

Table 4: This table presents the held out classification performance measures shown in Figure 3, when 100 out of 195 webpage labels are held out. The baseline AUC is 0.5. Random walk smoothing does slightly worse, while Tikhonov smoothing does better. Methods based on empirical stationary correlations bring larger improvements.

Sometimes we can make use of an asymmetry in the labels. For example, positive determinations, e.g. 1s, may have intrinsically higher confidence than negative determinations, −1s, and we can vary $\sigma_i$ to account for this. Similarly when one binary label in ±1 is relatively rare, we could use a value
other than 0 as our default guess.

In other applications there are covariate vectors $X_i \in \mathbb{R}^p$ measured at every node $i$ that we would like to take account of. It is standard in kriging to incorporate such a regression. Here as in other articles on semi-supervised learning over graphs, we emphasize the role of the graph structure, not the covariates.

Finally, it is not really necessary to have the same vector $\nu$ appear in both the variance model through $\sigma^2 VRV$ with $V = \text{diag}(\nu)$ and in the model for the mean through $\mu \nu$. In particular, for the models with $\nu = 0_n$, including two in Table 1, $\sigma^2 VRV + \Lambda$ reduces to $\Lambda$ and then our method for incorporating empirical correlations breaks down. For those, there is a simple procedure taking $1_n$ for the variance model and $0_n$ for the mean. Also, we could hybridize the Tikhonov and random walk models, using $\nu = 1_n$ from the former inside the regression model with the edge directionality respecting covariance of the latter.

8 Conclusion

We have shown that several recently developed semi-supervised learning methods for data on graphs can be expressed in terms of kriging. Those kriging models use implied correlations that derive from the graph structure but do not take account of sample correlations among the observed values.

Our proposed empirical stationary correlation model uses correlation patterns seen among the observed values to estimate a covariance matrix over the entire graph. In two numerical values we saw that using empirical correlations brought large improvements in performance. Even when there were large differences between the performance levels of different semi-supervised methods, the use of empirical correlations narrowed the gap. This reduces the penalty for the user who picks a suboptimal vector $\nu$.

We have not implemented our method on any large scale problems. Large scale presents two challenges. First, solving equations with an $n \times n$ matrix is expensive. Second, the number of correlation pairs $\hat{R}_{ij}$ to smooth is large. Reduced rank correlation matrices will mitigate the first problem. The second problem arises when the number $r$ of labeled cases is large. Large $r$ is much rarer than large $n$, and in any case can be mitigated by downsampling the correlation pairs before smoothing. In our examples covariance estimates derived from quite small numbers of observation pairs still performed well. We finish by pointing out that there are a good many smaller datasets to which semi-supervised learning on graphs may be applied.
Acknowledgments

This work was supported by grant number DMS-0604939 from the U.S. National Science Foundation.

References


