总结报告20

论文Multi-task sparse structure learning with Gaussian copula models

论文A framework for learning predictive structures from multiple tasks and unlabeled data

（2020.2.29）

一、contents

Gonçalves, A. R., Von Zuben, F. J., & Banerjee, A. (2016). **Multi-task sparse structure learning with Gaussian copula models**. Journal of Machine Learning Research, 17, 1–30.

Ando, R. K., & Zhang, T. (2005). **A framework for learning predictive structures from multiple tasks and unlabeled data**. Journal of Machine Learning Research, 6, 1817–1853.

二、论文1

1. 题目

Gonçalves, A. R., Von Zuben, F. J., & Banerjee, A. (2016). **Multi-task sparse structure learning with Gaussian copula models**. Journal of Machine Learning Research, 17, 1–30.

2. 内容

2.1 概述

提出的方法：MSSL——Multi-task Sparse Structure Learning

这篇论文给出了学习建立Ω的两种方法：



1）p-MSSL (from tasks parameters)——根据

2）r-MSSL (from residual error)——根据

抽象的目标函数：



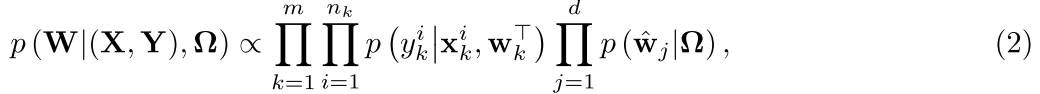
Note:

：inductive bias，captures the interaction between parameters wk and the relationship matrix Ω.

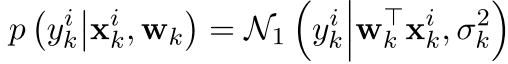
：sparsity inducing regularization

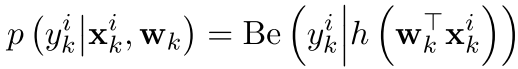
**pMSSL (from tasks parameters)：**

作者假设了W的行之间满足多元高斯分布的先验假设，那么后验假设可以写成：



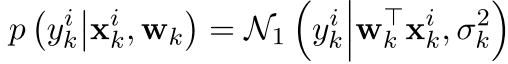
后面作者将具体为两种形式：

一种用在最小二乘回归，认为

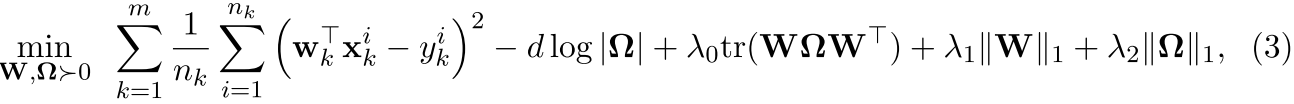
另一种用在逻辑回归（也即classification），认为

2.2 第一种——最小二乘回归问题

因为

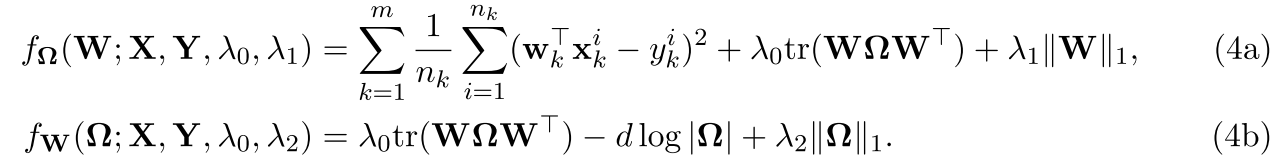


假设

Note：

enforcing sparsity on Ω will highlight the conditional independence among tasks parameters：任务参数间条件独立

由于（3）关于W,Ω非凸，但是单独对于W,Ω其中一个，（3）是凸的，所以可以交替优化，转化成：

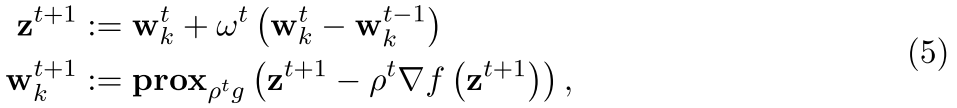


——Update for W：accelerated proximal gradient descent

近端算法的理论基础为：

其中，f(x)：凸光滑，g(x)：凸但是不光滑

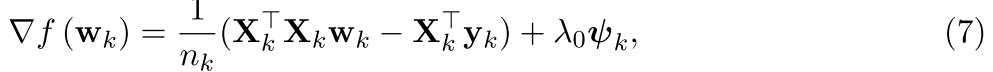
那么accelerated proximal gradient descent为：



其中，



论文（4a）中，

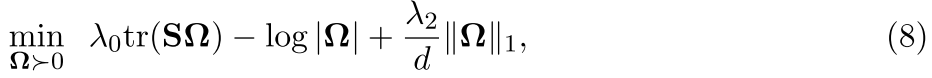


Where，

——Update for Ω：ADMM

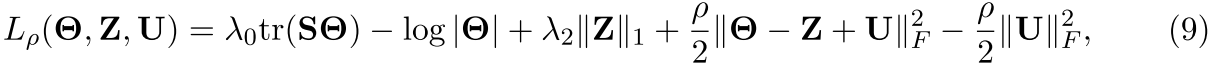
（4b）中Ω的更新就是著名的sparse inverse covariance selection problem

引入sample covariance matrix



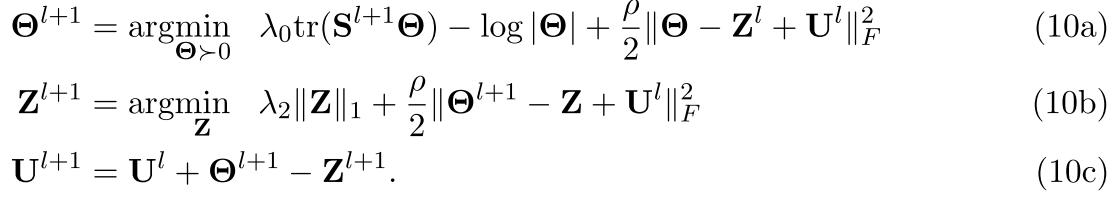
下面用ADMM来解决（8），

（8）的增广拉格朗日函数为：



其中，U为scaled dual variable

引入线性约束，ADMM的更新分解为下面三个迭代更新：



最终得到的结果为，L为迭代次数

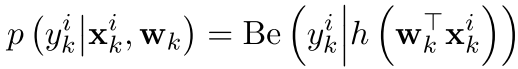
（10b）的更新可以简化为



得到的是闭型解（closed form？）

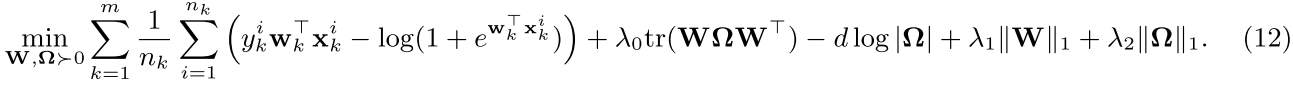
2.3 第二种——逻辑回归问题

因为

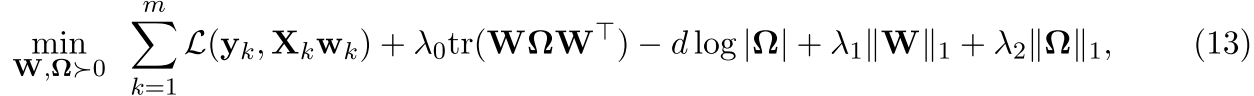


Note：Be（）为伯努利分布,h（）为sigmoid函数

所以，



一个更加general的目标函数，GLM（generalized linear model）：

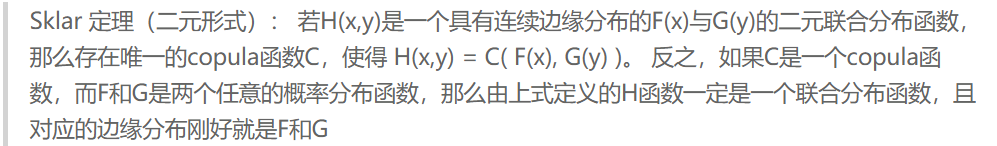


（12）计算起来和（3）是一样的方法。

2.4 引入了Gaussian Copula Model

（为什么要引入copula model？因为（4b）中的W被认为服从正态分布，在某些情况可能不合适）

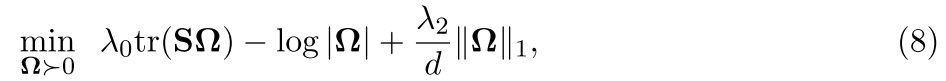
Gaussian Copula的理论基础：

1）Sklar定理：

简单的说，通过Sklar定理，我们可以将一个个多元随机变量的联合分布（joint distribution）分解为各个变量的边缘分布（marginal distribution）以及纯粹刻画他们之间关联性（dependence structure）的copula

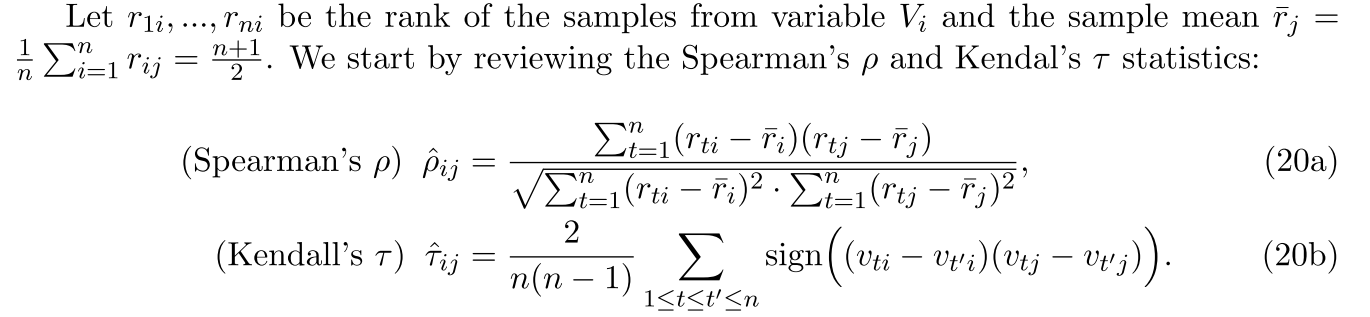
直接说结论：

通过Gaussian Copula，我们利用它来完成我们工作的一部分，那就是update Ω in（8），之后在接着用ADMM



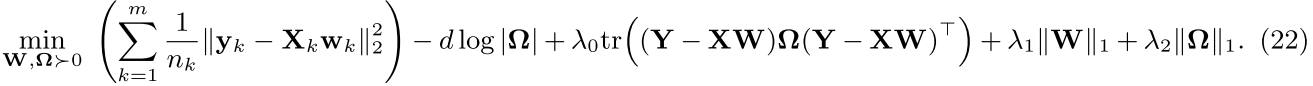




其中，

**r-MSSL (from residual error)——根据**





三、论文2

1. 题目

Ando, R. K., & Zhang, T. (2005). **A framework for learning predictive structures from multiple tasks and unlabeled data**. Journal of Machine Learning Research, 6, 1817–1853.

2. idea

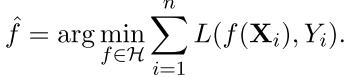
Learn **predictive structures on hypothesis spaces** (that is, what kind of classifiers have good predictive power) from multiple learning tasks

问题：什么叫data-manifold？

We want to learn some underlying **predictive functional structure**s (smooth function classes) that can characterize what good predictors are like. We call this problem structural learning. Our key idea is to learn such structures by considering multiple prediction problems simultaneously.

3. 内容

3.1 ERM



注意是f∈H，H=hypothesis space

With a fixed sample size, the smaller the hypothesis space H, the easier it is to learn the best predictor in H.

The error caused by learning the best predictor from finite sample is called the **estimation error**.

The smaller the hypothesis space H, the less accurate the best predictor in H becomes.

The error caused by using a restricted H is often referred to as the **approximation error**.

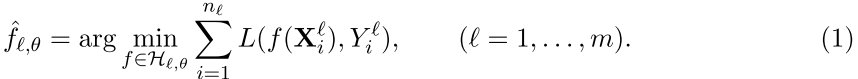
In supervised learning, one needs to select **the size of H** to balance the **trade-off between approximation error and estimation error**.

This is typically done through **model selection**, where we learn a set of predictors from a set of candidate hypothesis spaces Hθ, and then pick the best choice on a validation set.

3.2 符号说明

m个learning problems indexed by ，每个有个examples

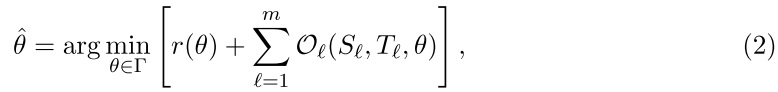
给定fixed structural parameter θ



The θ parameter encodes our assumption of what a good predictor should be like.

3.3 structural learning

假设，并且认为通常，if two points are close in this intrinsic distance（内在的距离）, then the values that a good predictor produces at these points are also likely to be similar.

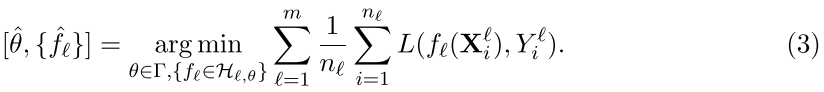


Note:

r（θ）为正则项，

另外可以让

联合估计：



3.4 linear predictors

原空间可能没有线性关系，利用映射映射到高维，可能就有线性关系了，所以，这样的话，找f实际就是找，因为。

假设



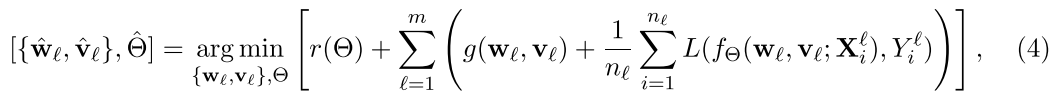
前者 a known high-dimensional feature map

后者 a parameterized low-dimensional feature map

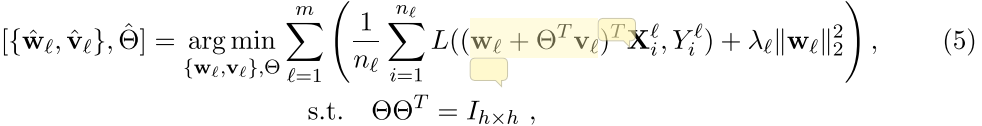
为了简化令，于是



目标函数变成，

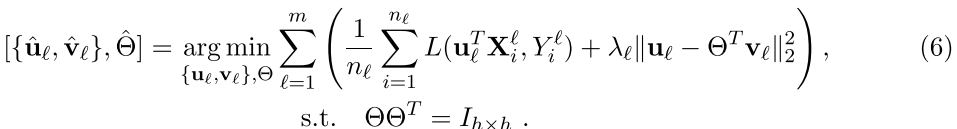


再简化，令，



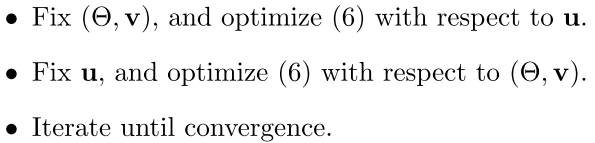
引入，

得到最终问题，

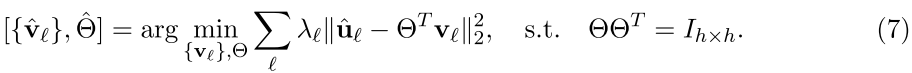


3.5 ASO

ASO算法：



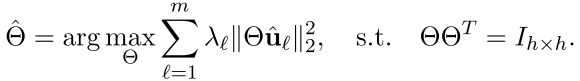
给定u，计算（Θ，v）



给定Θ，计算v



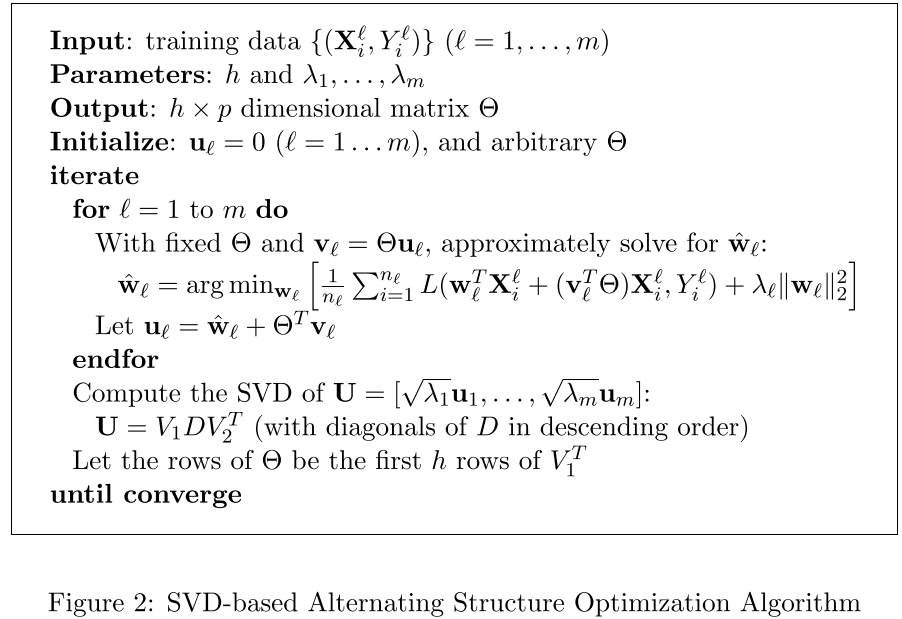
给定，计算Θ



令 ，上式变为



具体见下：



Note：SVD—ASO和PCA不同在于，ASO是对predictor space降维而不是data space.

另外，也可以对Feature（component of x）进行分组，而不是instance.

四、问题

1.data-manifold是什么意思？

2.闭型解（closed form）是什么意思？