Incorporation of Biological Pathway Knowledge in the Construction of Priors for Optimal Bayesian Classification

Mohammad Shahrokh Esfahani

Department of Electrical and Computer Engineering
The Center for Bioinformatics and Genomic Systems Engineering
Texas A&M University

February 22, 2014
Small sample is commonplace in proteomic/genomic classification: data is rare, expensive, or difficult to acquire

On the other hand: data-driven methods work well only with a large amount of data.

And, model-free classification is virtually impossible.

**Why?** Without assumptions regarding the model, we have no guarantees about the accuracy of any prediction.

Mostly, leading to poor supervised predictions and predictive power estimates (error estimation).
A promising approach to alleviate the problem is the use of *prior knowledge*.

A model-based view of quantifying our prior knowledge is to construct prior probability. Prior probability assigns each model a probability: It is compressing our uncertainty about the model of the underlying system; in classification, uncertainty is about class-conditional densities $f(x|y)$.

Using “prior probability” in the design of Optimal Bayesian Classifiers (OBC) has been recently introduced by Dalton and Dougherty, 2012.
Two components of any Bayesian analysis: prior probability, and data!

Prior: Consider only $f(x, y)$ consistent with scientific knowledge

Posterior: more weight on models consistent with observations

Where do these “prior probabilities” come from? Is this enough to choose them “subjectively?” Are they really important?
Motivation IV

- How much a “contaminated prior” can affect the performance?

Example of multivariate Gaussian with Normal-inverse-Wishart prior

Deviation the prior’s center from the true probability can deteriorate the performance
Focusing on the most relevant type of prior knowledge in phenotype classification: biological pathways.

The system here is a combination of interactions and regulatory signals between components, being gene or protein.

Example of prior knowledge: pathways

What is the missing step?

prior knowledge $\rightarrow$ prior probability
Our Work II

- First step: *transforming the prior knowledge (pathways) to “testable information”*

- Second step: *mapping the “testable information” to the hyperparameter space using some model*

**Novelty**

*The main novelty lies in that we construct priors using some part of the sample data.*
Prior Properties

The prior distribution family, \( \Pi \) (from which one is selected) should be analytically tractable in: It should be...

- Reasonably easy to determine the posterior distribution given a “sample”.
- Possible to express conveniently the expectations of some desired functions.
- Closed, the posterior belongs to the family which the prior comes from (i.e. conjugate priors).
- Rich, so that there will exist a member of \( \Pi \) capable of expressing any prior information.
- Parametrized in a manner which can be readily interpreted in relation to prior information.
From Biological Pathways to Testable Information
Suppose that the model is parameterized by $\theta$: $\pi \rightarrow \pi(\theta)$, set of models $\mathcal{M}(\theta)$.

### Simplified wiring of colon cancer-related pathways $\mathcal{G}$

- APS $\mathcal{G}_a$:
  - $\text{EGF} \rightarrow \text{RAS}$
  - $\text{HGF} \rightarrow \text{RAS}$
  - $\text{HGF} \rightarrow \text{PIK3CA}$
  - $\text{IL6} \rightarrow \text{STAT3}$
  - $\text{IL6} \rightarrow \text{RAS}$

- RPS $\mathcal{G}_r$:
  - $\text{PIK3CA} \rightarrow \text{STAT3}$
  - $\text{TSC1/TSC2} \rightarrow \text{mTOR}$
  - $\text{SPRY4} \rightarrow \text{PKC}$

### Pairwise regulations to testable information

- APS: $E_{\mathcal{M}(\theta)}[\Pr(\text{RAS} = \text{up-reg} | \text{EGF} = \text{up-reg})] \geq 1 - \xi^a$ for some small $\xi^a > 0$

- RPS: $E_{\mathcal{M}(\theta)}[\Pr(\text{STAT3} = \text{down-reg} | \text{PIK3CA} = \text{up-reg})] \geq 1 - \xi^r$ for some small $\xi^r > 0$
Biological Pathways– Soft Constraints II

... pathways

regulatory sets $\overline{R}_x$

$\overline{R}_{\text{EGF}} = \{\text{RAS, STAT3}\}$
$\overline{R}_{\text{HGF}} = \{\text{RAS, PIK3CA}\}$
$\overline{R}_{\text{IL6}} = \{\text{RAS, STAT3, PKC}\}$
$\overline{R}_{\text{RAS}} = \{\text{PIK3CA, MEK1/2}\}$

$C = \{\forall x(i) \in G : R_{x(i)} \neq \emptyset\}$

regulatory set to testable info via conditional entropy

discrete case: $E_{M(\theta)}[H[\text{EGF}|R_{\text{EGF}}]] \leq \xi^{\text{reg}}$ for some small $\xi^{\text{reg}} > 0$

differential entropy: $E_{M(\theta)}[H[\text{EGF}|R_{\text{EGF}}]] \leq \xi^{\text{reg}}$
Regularized Expected Mean-log-likelihood Priors: Combining Data with Pathways
data-related function

mean log-likelihood:
\[ \ell_{np}(\theta) := \frac{1}{np} \ell(\theta; S_{np}^{prior}) = \frac{1}{np} \sum_{i=1}^{np} \log f(X_i|\theta); X_i \in S_{np}^{prior} \]

model selection view: \( \ell_{np}(\theta) \) is a plug-in estimate of
\[ \int_{x \in \mathcal{X}} f(x|\theta_{true}) \log f(x|\theta) dx = -D_{KL}(f(x|\theta_{true}) || f(x|\theta)) + g(\theta_{true}) \]

prior-expected mean log-likelihood (Bayesian perspective):
\[ E_M(\theta)[\ell_{np}(\theta)] \]
interpretation: uncertainty-class-averaged similarity with the true model
Regularized Expected Mean-log-likelihood Priors II

Regularized Expected Mean log-Likelihood (REML) prior

\[
\min_{\pi(\theta) \in \Pi, \xi} - (1 - \lambda_1 - \lambda_2) E_\theta \left[ \ell_{np}(\theta) \right] + \lambda_1 \sum_{i=1}^{|C|} \xi_i^{reg} + \lambda_2 \left[ \sum_{(i,j) \in G_a} \xi_i^a + \sum_{(i,j) \in G_r} \xi_i^r \right]
\]

subject to the pathway-based constraints

\[
\begin{align*}
E_\theta \left[ H_\theta \left[ x(i) \mid R_{x(i)} \right] \right] & \leq \xi_i^{reg}; \forall x(i) \in C \\
E_\theta \left[ \Pr(x(j) = UR \mid x(i) = UR) \right] & \geq 1 - \xi_i^{a}; x(i) \rightarrow x(j) \in G_a \\
E_\theta \left[ \Pr(x(j) = DR \mid x(i) = UR) \right] & \geq 1 - \xi_i^{r}; x(i) \rightarrow x(j) \in G_r \\
\end{align*}
\]

Gaussian-relaxation to correlation coeffs

\[
\begin{align*}
E_\theta \left[ \rho_{x(i),x(j)} \right] & \geq 1 - \xi_i^{a}; x(i) \rightarrow x(j) \in G_a \\
E_\theta \left[ \rho_{x(i),x(j)} \right] & \leq -1 + \xi_i^{r}; x(i) \rightarrow x(j) \in G_r \\
\end{align*}
\]
Overall Strategy

Motivation

Path. Bayesian Quant.

REML for $\mathcal{N}\mathcal{W}^{-1}$ Prior

Numerical Exp.

Summary

Prior Construction

Regularized Expected Mean-log-likelihood Priors: Combining Data with Pathways

- Biological pathways
- REML prior construction
- Sampling from $S_n$ to generate prior constructing sample points
- Exclude prior constructing points ($S_n \setminus S_{np}$)
- OBC training
- $\psi_{OBC}$
REML for the Multivariate Gaussian with Normal-inverse-Wishart Prior
Motivation

Prior Construction

Path. Bayesian Quant.

REML for $\mathcal{N}\mathcal{W}^{-1}$ Prior

Numerical Exp.

Summary

Q?

**Multivariate Gaussian**

Variables are jointly Gaussian, i.e. $x \sim \mathcal{N}(\mu, \Lambda^{-1})$. **Uncertainty about the mean and covariance matrix.**

**Colon cancer pathway**

Data collection

Prior over Gaussian distributions

Posterior over Gaussian distributions

Some portion of the data

True $f(x, y)$

Prior peak

Posterior peak

True $f(x, y)$

Mohammad Shahrokh Esfahani

REML for the Multivariate Gaussian with Normal-inverse-Wishart Prior
Normal-inverse-Wishart I

consider a Gaussian model: \( \theta = [\mu, \Lambda] \)
then, estimate hyperparameters when

\[
\pi(\theta) \in \Pi = \{ \mathcal{N}(m, \nu, W, \kappa) : m \in \mathbb{R}^p, W > 0 \}
\]

the Normal-Wishart has four parameters, \([m_{p \times 1}, \nu, W_{p \times p}, \kappa] :\)

\[
\mu|\Lambda \sim \mathcal{N}(m, (\nu \Lambda)^{-1}) \\
\Lambda = \Sigma^{-1} \sim \mathcal{W}(W, \kappa) \propto |W|^{-\kappa/2} |\Lambda|^{(\kappa-p-1)/2} \exp\{-\frac{1}{2}\text{tr}(W^{-1}\Lambda)\},
\]

assume known \( \kappa \) and \( \nu \): chosen by the practitioner to represent the strength of his/her conviction

break the general non-convex problem into two consecutive convex progs... easy to show that \( \hat{m} = \mu_m \); yet the trick: writing everything with \( \lambda_2 = 0 \) as a function of precision matrix and its components, instead of cov. matrix itself
Normal-inverse-Wishart II

\[
\Lambda = \begin{bmatrix}
\Lambda_{R_x} & \Lambda_{12} & \Lambda_{13} \\
\Lambda_{21} & \Lambda_x & \Lambda_{23} \\
\Lambda_{31} & \Lambda_{32} & \Lambda_{33}
\end{bmatrix} \quad ; \quad W = \begin{bmatrix}
W_{R_x} & W_{12} & W_{13} \\
W_{21} & W_x & W_{23} \\
W_{31} & W_{32} & W_{33}
\end{bmatrix}.
\]

reml with \( \lambda_2 = 0 \); \( CP_1(\kappa) \):

\[
\min_{W > 0, \xi_i^{reg}} \quad -\frac{1}{2}(1 - \lambda_1) \left[ \log |W| - \kappa \text{tr}(WW) \right] + \lambda_1 \sum_{i=1}^{C} \xi_i^{reg} \\
\text{subject to} \quad -\log \sqrt{|W_{x(i)}|} - \psi \left( \frac{\kappa - (p - |R_{x(i)}| - 1)}{2} \right) \leq \xi_i^{reg},
\]

where: \( \sqrt{|W_{x(i)}|} := W_{x(i)} - W_{x(i),g\bar{R}_{x(i)}W^{-1}g\bar{R}_{x(i)}}W^T_{x(i),g\bar{R}_{x(i)}}. \)

**Lemma**

*The programming, \( CP_1(\kappa) \) is convex in \( W \) and satisfies the Slater’s condition.*
Normal-inverse-Wishart III

Approx. of corr. coeffs using outcome of CP\(_1(\kappa)\): \(W^* = \Psi^* - 1\)

Known result: if \([\sigma_{ij}]_{p \times p} \sim W^{-1}(\Psi, \kappa)\), \(E[\sigma_{ij}] = \frac{1}{k-p-1} \psi_{ij}\), 
\(i, j \in \{1, \ldots, p\}\), approximate

\[
E[\rho_{ij}] = E \left[ \frac{\sigma_{ij}}{\sqrt{\sigma_{ii} \sigma_{jj}}} \right] \approx \frac{E[\sigma_{ij}]}{1 - k-p-1 \sqrt{\psi_{ii}^* \psi_{jj}^*}} = \frac{\psi_{ij}}{\sqrt{\psi_{ii}^* \psi_{jj}^*}}.
\]

Part II of reml: \(CP_2(\Psi^*)\):

\[
\min_{\Psi > 0, \xi_{i,j}^a, \xi_{i,j}^r \geq 0} (1 - \lambda_2) ||\Psi - \Psi^*||_F^2 - \lambda_2 \left[ \sum_{(i,j) \in G_a} \xi_{i,j}^a + \sum_{(i,j) \in G_r} \xi_{i,j}^r \right]
\]

Subject to

\[
\begin{cases}
1 - \xi_{i,j}^a \leq \frac{\psi_{ij}}{\sqrt{\psi_{ii}^* \psi_{jj}^*}} \leq 1; x(i) \rightarrow x(j) \in G_a \\
1 - \xi_{i,j}^r \leq \frac{-\psi_{ij}}{\sqrt{\psi_{ii}^* \psi_{jj}^*}} \leq 1; x(i) \rightarrow x(j) \in G_r \\
\psi_{ij} = \psi_{ji}
\end{cases}
\]

The programming \(CP_2(\Psi^*)\) is convex.
Numerical Experiments

Simulations on both synthetically generated and real pathways
Gaussian feature-label distributions (classification problem)

1. fix true parameterization for two classes: $[\mu_y^{true}, \Sigma_y^{true}], y \in \{0, 1\}$.
2. generate two sets of pathways, $G_y, y \in \{0, 1\}$.
3. take observations from $N(\mu_y^{true}, \Sigma_y^{true})$ to generate $S_n$.
4. randomly choose $n_p$ points from $S_n$ for prior construction, i.e., $S_{n_p}^{prior}$, and the rest $S_{n_t}^{train}$ for training.
5. use $S_{n_p}^{prior}$ and $G_y$ to construct the prior $\pi_y^{reml}, y \in \{0, 1\}$, by reml (CP$_1(\kappa)$ and CP$_2(\Psi^*)$).
6. optimally combine the priors, $\pi_y^{reml}, y \in \{0, 1\}$, and $S_{n_t}^{train}$ to build the obc, $\psi_{obc,n_t}^{n_p}$.
7. approximate the expected error rate of the designed classifier.
use blocked covariance matrix structure proposed for modeling the gene expression microarrays [Hua, et. al, 2005]

\[
\Sigma = \begin{bmatrix}
B_1 & C & C \\
C & B_2 & C \\
C & C & B_3 \\
\end{bmatrix}, \quad B_i = \begin{bmatrix}
\sigma^2 & \rho_i \sigma^2 & \rho_i \sigma^2 \\
\rho_i \sigma^2 & \sigma^2 & \rho_i \sigma^2 \\
\rho_i \sigma^2 & \rho_i \sigma^2 & \sigma^2 \\
\end{bmatrix}, \quad B_i = \begin{bmatrix}
\sigma^2 & \rho_i \sigma^2 & \rho_i \sigma^2 \\
\rho_i \sigma^2 & \sigma^2 & \rho_i \sigma^2 \\
\rho_i \sigma^2 & \rho_i \sigma^2 & \sigma^2 \\
\end{bmatrix}
\]

| Class | \( \Sigma_{y}^{true} \) | \( \mu_{y}^{true} \) | \( |\mathcal{E}| \) | \( |\mathcal{O}| \) | \( \nu_{y} \) | \( M \) |
|-------|-----------------|----------------|--------|--------|---------|--------|
| 0     | \( \rho_1 = \rho_3 = 0.3 \) \( \rho_2 = -0.3, \rho_c = 0.1 \) | C1: 0.31_p | 50     | 100    | \( n_0^p \) | 15000  |
| 1     | \( 2 \Sigma_{1-y}^{true} \) | C1: \(-0.31_p\) | 50     | 100    | \( n_1^p \) | 15000  |
Synthetic Pathways and Data I

C1: $c = 0.5; \epsilon_{\text{Bayes}} = 0.167; x$-axis: $\frac{n^p_0 + n^p_1}{n}$ (\%); $\kappa_y = 2p + n^p_y$; left: $n = 50$; right: $n = 70$
Synthetic Pathways and Data II

C2: \( c = 0.5; \epsilon_{\text{Bayes}} = 0.091; \) x-axis: \( \frac{n_0^p + n_1^p}{n} \) (\%); \( \kappa_y = 2p + n_y^p; \) left: \( n = 50; \) right: \( n = 70 \)
linear dependencies

\[ x = [\text{EGF, HGF, IL6, Ras, PIK3CA, STAT3, TSC1/TSC2, mTORC1, SPYR4, PKC, MEK1/2}] \]

\[ [\text{EGF, HGF, IL6}] \sim \mathcal{N}([\mu_0^{true}] [x(1) \ x(2) \ x(3)]; [\Sigma_0^{true}] [x(1) \ x(2) \ x(3)]) \]

\[ x(i) = a_i^T x_{i-1} + z_i; i = 4, 5, ..., 11; \ z_i \sim N(0, \sigma_i^2) \]

for example: \( \text{Ras} = a_4(1)\text{EGF} + a_4(2)\text{HGF} + a_4(3)\text{IL6} + z_4 \)

for \( y = 1 \), we assume \( \text{TSC1/TSC2} \) (tumor suppressor complex):
\( \text{TSC1/TSC2} = z_7 \)

<table>
<thead>
<tr>
<th>Class ( y )</th>
<th>( [\Sigma_y^{true}] [x(1) \ x(2) \ x(3)] )</th>
<th>( \mu_y^{true} )</th>
<th>Noise variance</th>
</tr>
</thead>
</table>
| 0             | \[
\begin{bmatrix}
1 & 0.2 & 0.2 \\
0.2 & 1 & 0.2 \\
0.2 & 0.2 & 1
\end{bmatrix}
\] | \( 0.31_p \) | \( \sigma_i^2 = 0.2, \ i = 1, ..., 8 \) |
| 1             | \[
\begin{bmatrix}
2 & 0.4 & 0.4 \\
0.4 & 2 & 0.4 \\
0.4 & 0.4 & 2
\end{bmatrix}
\] | \( -0.31_p \) | \( \sigma_i^2 = 0.05, \ i \neq 7 \) |
Real Pathways- Synthetic Data III

\[ \epsilon_{\text{Bayes}} = 0.132; \text{x-axis: } \frac{n_0^p + n_1^p}{n} \text{ (\%); } \kappa_y = 2p + n_y^p; \text{ left: } n = 50; \text{ right: } n = 70 \]
Summary I

Our work:

- Combined two sources of information: observed sample data (microarray or RNA-seq gene expression data) and biological signaling pathways via REML framework
- For the Gaussian case, the REML framework provides two convex optimization problems guaranteed to converge

In general:

- There is a great potential to enhance phenotype classification accuracy using the existing pathways
- "subjectivity" can be mitigated in Bayesian approaches using sophisticated methods for prior construction
- REML method is the first attempt towards incorporation of prior knowledge for prior construction in genomics
What about \textbf{prior construction} for Bayesian classification using Next-generation Sequencing (NGS) data?

- This is all about modeling.
- The priors must be constructed, using some model for the NGS data, e.g. multivariate Poisson model for the read counts:
  - Hyperparameters are the parameters of the distribution governing the multivariate Poisson parameter vector.

Chain: Pathways $\rightarrow$ Testable Information $\rightarrow$ Objective function of hyperparameters.
Ronald A. Fisher, 1925:

“Little experience is sufficient to show that the traditional machinery of statistical processes is wholly unsuited to the needs of practical research. Not only does it take a canon to shoot a sparrow, but it misses the sparrow!... Only by systematically tackling small sample problems on their merits does it seem possible to apply accurate tests to practical data.”

Thank you...
Any Question?