Patient Representation

Derek Modzelewski

October 19, 2017

1 Weighting Patients

Not all patients had the same tissues sampled. Furthermore, not all patients have the same *number* of tissues sampled. Thus, we are more confident about the learned representations for some patients (the ones with more samples) than for other patients (the ones with fewer samples). This factors into our prediction of the tissue transform matrices and tissue centers, which then affect the prediction of patient representations.

Instead of solving for the optimal weights directly, we use the weights generated from a related problem which is significantly easier to work with. It is our belief that the weighting scheme will be optimal for our patient representation problem as well.

Theorem 1.1. *Given many estimations $\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_n$ of parameter $\theta_0 \in \mathbb{R}^d$, where*

$$\hat{\theta}. \equiv (\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_n) = \theta_0 \mathbf{1}^T + B + u + r$$

$$\mathbb{E}[u] = \mathbb{E}[r] = \mathbf{0}$$

where B is the bias, u is the "uncertainty" in the bias and has mean 0 and covariance matrix ϵ, and r is the "random error" and has mean 0 and covariance DVD where D is a diagonal matrix where $D_{i,i} = 1/\sqrt{n_i}$ $\forall i \in [n]$, and u and r are independent
Then the estimation $\tilde{\theta}$ of θ_0 which minimizes squared loss is given by $w^T\hat{\theta}.$ where

$$w = k(2B^TB + \epsilon + DVD)^{-1}\mathbf{1}$$

for suitable $k \in \mathbb{R}$ s.t. $w^T\mathbf{1} = 1$

Proof. any estimation $\delta(\hat{\theta}.)$ of θ_0 using can be expressed as $\delta(\hat{\theta}.) = \hat{\theta}.w + v$ for some $w \in \mathbb{R}^n$ *s.t.* $\mathbf{1}^Tw = 1$, some $v \in \mathbb{R}^d$, for given $\hat{\theta}.$

$$\begin{aligned}
\mathbb{E}\big[||\delta(\hat{\theta}.) - \theta_0||^2\big] &= \mathbb{E}\big[||\hat{\theta}.w + v - \theta_0||^2\big] \\
&= \mathbb{E}\big[||\theta_0\mathbf{1}w + Bw + uw + rw + v - \theta_0||^2\big] \\
&= \mathbb{E}\big[||\theta_0 + Bw + uw + rw + v - \theta_0||^2\big] \\
&= \mathbb{E}\big[||Bw + uw + rw + v||^2\big] \\
&= \mathbb{E}\big[||(B + u + r)w + v||^2\big] \\
&= \mathbb{E}\big[w^T(B + u + r)^T(B + u + r)w + 2v^T(B + u + r)w + v^Tv\big] \\
&= \mathbb{E}\big[w^T(B + u + r)^T(B + u + r)w\big] + 2v^T\mathbb{E}\big[B + u + r\big]w + v^Tv \\
&= w^T\mathbb{E}\big[(B + u + r)^T(B + u + r)\big]w + 2v^TBw + v^Tv \\
&= w^T\mathbb{E}\big[B^TB + u^Tu + r^Tr\big]w + 2v^TBw + v^Tv \\
&= w^T\mathbb{E}(B^TB + \epsilon + DVD)w + 2v^TBw + v^Tv
\end{aligned}$$

At minimum, the gradient with respect to w and with respect to v will be normal to the feasible spaces of w and u, respectively.

v has full rank feasible region (no constraints), so only the 0 vector is normal to its feasible region:

$$\begin{aligned}\Rightarrow 0 &= \nabla_v\Big(\mathbb{E}\big[||\delta(\hat{\theta}.) - \theta_0||^2\big]\Big)\\ &= \nabla_v\Big(w^T(B^TB + \epsilon + DVD)w + 2v^TBw + v^Tv\Big)\\ &= \nabla_v\Big(2v^TBw + v^Tv\Big)\\ &= 2Bw + 2v\end{aligned}$$

$$\Rightarrow v = Bw$$

For w, our only constraint is that $\mathbf{1}^Tw = 1 \therefore$ forall feasible w, all vectors d normal to the feasible region can be expressed $d = k\mathbf{1}$ for some $k \in \mathbb{R}$

$$\begin{aligned}\Rightarrow k\mathbf{1} &= \nabla_w\Big(\mathbb{E}\big[||\delta(\hat{\theta}.) - \theta_0||^2\big]\Big)\\ &= \nabla_w\Big(w^T(B^TB + \epsilon + DVD)w + 2v^TBw + v^Tv\Big)\\ &= \nabla_w\Big(w^T(B^TB + \epsilon + DVD)w + 2v^TBw\Big)\\ &= 2(B^TB + \epsilon + DVD)w + (2v^TB)^T\\ &= 2(B^TB + \epsilon + DVD)w + 2B^Tv\\ &= 2(B^TB + \epsilon + DVD)w + 2B^TBw\\ &= 2(2B^TB + \epsilon + DVD)w\end{aligned}$$

$$\Rightarrow w = k(2B^TB + \epsilon + DVD)^{-1}\mathbf{1}$$

for appropriate k

2 Weighting Patients

Not all patients had the same tissues sampled. Furthermore, not all patients have the same *number* of tissues sampled. Thus, we are more confident about the learned representations for some patients (the ones with more samples) than for other patients (the ones with fewer samples). This factors into our prediction of the tissue transform matrices and tissue centers, which then affect the prediction of patient representations.

Instead of solving for the optimal weights directly, we use the weights generated from a related problem which is significantly easier to work with. It is our belief that the weighting scheme will be optimal for our patient representation problem as well.

Theorem 2.1. *Given many estimations $\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_n$ of parameter $\theta_0 \in \mathbb{R}^d$, where*

$$\hat{\theta}. \equiv (\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_n) = \theta_0\mathbf{1}^T + r$$

Then the estimation $\tilde{\theta}$ of θ_0 which minimizes squared loss is given by $\hat{\theta}.w + v$ where

$$w = k(\mathbb{E}[r^Tr] - \mathbb{E}[r]^T\mathbb{E}[r])^{-1}\mathbf{1}$$

for suitable $k \in \mathbb{R}$ s.t. $w^T\mathbf{1} = 1$, and

$$v = -\mathbb{E}[r]w$$

Proof. any estimation $\delta(\hat{\theta}.)$ of θ_0 using can be expressed as $\delta(\hat{\theta}.) = \hat{\theta}.w + v$ for some $w \in \mathbb{R}^n$ *s.t.* $\mathbf{1}^T w = 1$, some $v \in \mathbb{R}^d$, for given $\hat{\theta}$.

$$\begin{aligned}
\mathbb{E}\big[||\delta(\hat{\theta}.) - \theta_0||^2\big] &= \mathbb{E}\big[||\hat{\theta}.w + v - \theta_0||^2\big] \\
&= \mathbb{E}\big[||\theta_0 \mathbf{1}^T w + rw + v - \theta_0||^2\big] \\
&= \mathbb{E}\big[||rw + v||^2\big] \\
&= \mathbb{E}\big[w^T r^T rw + 2v^T rw + v^T v\big] \\
&= w^T \mathbb{E}[r^T r]w + 2v^T \mathbb{E}[r]w + v^T v
\end{aligned}$$

At minimum, the gradient with respect to w and with respect to v will be normal to the feasible spaces of w and u, respectively.

v has full rank feasible region (no constraints), so only the 0 vector is normal to its feasible region:

$$\begin{aligned}
\Rightarrow 0 &= \nabla_v \Big(\mathbb{E}\big[||\delta(\hat{\theta}.) - \theta_0||^2\big]\Big) \\
&= \nabla_v \Big(w^T \mathbb{E}[r^T r]w + 2v^T \mathbb{E}[r]w + v^T v\Big) \\
&= \nabla_v \Big(2v^T \mathbb{E}[r]w + v^T v\Big) \\
&= 2\mathbb{E}[r]w + 2v
\end{aligned}$$

$$\Rightarrow v = -\mathbb{E}[r]w$$

For w, our only constraint is that $\mathbf{1}^T w = 1 \therefore$ forall feasible w, all vectors d normal to the feasible region can be expressed $d = k\mathbf{1}$ for some $k \in \mathbb{R}$

$$\begin{aligned}
\Rightarrow k\mathbf{1} &= \nabla_w \Big(\mathbb{E}\big[||\delta(\hat{\theta}.) - \theta_0||^2\big]\Big) \\
&= \nabla_w \Big(w^T \mathbb{E}[r^T r]w + 2v^T \mathbb{E}[r]w + v^T v\Big) \\
&= \nabla_w \Big(w^T \mathbb{E}[r^T r]w + 2v^T \mathbb{E}[r]w\Big) \\
&= 2\mathbb{E}[r^T r]w + \big(2v^T \mathbb{E}[r]\big)^T \\
&= 2\mathbb{E}[r^T r]w + 2\mathbb{E}[r]^T v \\
&= 2\mathbb{E}[r^T r]w - 2\mathbb{E}[r]^T \mathbb{E}[r]w \\
&= 2\big(\mathbb{E}[r^T r] - \mathbb{E}[r]^T \mathbb{E}[r]\big)w \\
\Rightarrow w &= k(\mathbb{E}[r^T r] - \mathbb{E}[r]^T \mathbb{E}[r])^{-1}\mathbf{1}
\end{aligned}$$

for appropriate k

Corollary 2.1.1. *If we restrict our estimation to be in the convex hull of* $\{\hat{\theta}_1, ..., \hat{\theta}_n\}$*, the optimal solution is given by* $\hat{\theta}.w$ *where*

$$w = k\mathbb{E}[r^T r]^{-1}\mathbf{1}$$

for suitable k *as before.*

Proof. Any point in the convex hull of $\{\hat{\theta}_1, ..., \hat{\theta}_n\}$ may be expressed as $\hat{\theta}.w$, which is equivalent to setting v to $\mathbf{0}$. In the previous proof, we arrived at

$$\begin{aligned}
...k\mathbf{1} &= 2\mathbb{E}[r^T r]w + 2\mathbb{E}[r]^T v \\
&= 2\mathbb{E}[r^T r]w \qquad\qquad \because v = \mathbf{0}
\end{aligned}$$

$$\Rightarrow w = k\mathbb{E}[r^T r]^{-1}\mathbf{1}$$

Corollary 2.1.2. *In the special case that $r = u + v$ where*

$$\mathbb{E}[u] = \mathbb{E}[v] = \mathbf{0}$$

and $\mathbb{E}[u^T u] = \phi I$ and $\mathbb{E}[v^T v]$ is diagonal where $\mathbb{E}[v^T v]_{i,i} = \gamma/n_i \ \forall i \in [n]$, and we are restricted to estimations in the convex hull of $\{\hat{\theta}_1, ..., \hat{\theta}_n\}$, the optimal estimator, $\tilde{\theta}$, is

$$\tilde{\theta} = \frac{\sum_{i=1}^{n} \frac{n_i}{n_i + \gamma/\phi} \hat{\theta}_i}{\sum_{i=1}^{n} \frac{n_i}{n_i + \gamma/\phi}}$$

Remark 1. *This is equivalent to placing a $n_i/(n_i + \gamma/\phi)$ weight to each estimator θ_i*

Proof. ...