Supplemental Methods: Mathematics of CoGAPS

August 8, 2018

1 Atomic prior

CoGAPS decomposes a matrix $\mathbf{D}$ of G genes (rows) and S samples (columns) into two matrices $\mathbf{A} \in \mathbb{R}^{G \times k}$ and $\mathbf{P} \in \mathbb{R}^{k \times S}$ using the model:

$$p(\mathbf{A}, \mathbf{P} | \mathbf{D}, \mathbf{\Sigma}) \propto p(\mathbf{D} | \mathbf{A}, \mathbf{P}, \mathbf{\Sigma}) \, p(\mathbf{A}) \, p(\mathbf{P}), \tag{1}$$

where the elements of $\mathbf{\Sigma}$ represent the corresponding standard deviation of each element in the matrix $\mathbf{D}$. Determining the optimal value of k remains an open problem for latent space detection. The CoGAPS model assumes each element of $\mathbf{D}$ is i.i.d. with $p(D_{i,j} | A_{i,\cdot}, P_{\cdot,j}, \Sigma_{i,j})$ a normal distribution with mean $\mu_{i,j} = A_{i,\cdot} \times P_{\cdot,j}$ and variance $\sigma^2_{i,j}$.

In the case of sequencing data, $D_{i,j}$ is log transformed counts. In cases with replicates, $D_{i,j}$ can be replaced with the mean log transformed read counts and standard deviation can be computed across these replicates. In cases without replicates, the standard deviation is assumed to be 10% of the signal in $\mathbf{D}$ with a minimum value of 0.1.

CoGAPS uses an atomic prior [3] for the $\mathbf{A}$ and $\mathbf{P}$ matrices based upon previous work in Bayesian non-negative NMF for microarrays [2]. The atomic prior [3] is similar to spike and slab model [1], in which only a subset of model parameters are non-zero and those that are have a value distributed according to some continuous distribution with non-negative support. As a result, this model results in a ℓ_0 sparsity constraint on these matrices with other constraints depending on the distribution used to model non-zero values in these matrices. The atomic prior models each non-zero matrix element of $\mathbf{A}$ or $\mathbf{P}$ with a gamma distribution. The rate λ^A and λ^P of this distribution is a parameter that is fixed for every matrix element in $\mathbf{A}$ and $\mathbf{P}$, respectively. The shape of the gamma prior for each matrix element is a separate hyperparameter ($\alpha^A_{i,k}$ for each element of $\mathbf{A}$ and $\alpha^P_{k,j}$ for each element of $\mathbf{P}$), modeled as a Poisson distribution with a fixed parameter α for each matrix element. Zero values for $\alpha^A_{i,k}$ or $\alpha^P_{k,j}$ correspond to $A_{i,j} = 0$ and $P_{k,j} = 0$, modeling the subset of model parameters that are zero.

The expectation of the Gamma distribution is proportional to the sampled values of $\alpha^A_{i,k}$ or $\alpha^P_{k,j}$, introducing a further sparsity constraint on the magnitude of the matrix elements when these values are small. In contrast to standard spike and slab models, the atomic prior also models smoothness by encoding a correlation structure between matrix elements in $\mathbf{A}$ and $\mathbf{P}$ during the sampling steps.

Recall that $A_{i,j} \sim \Gamma\left(\alpha^A_{i,k}, \lambda^A\right)$ is equivalent to the sum of $\alpha^A_{i,k}$ independent, exponentially distributed random variables with rate parameter λ^A and similarly for $P_{k,j}$. Instead of directly sampling from the Gamma or Poisson distributions, the proposal distribution in the atomic prior updates a single, exponentially distributed random variable $x^A_{i,k,l}$ for A and $x^P_{k,j,m}$ for P at each step described in Section 2. The advantage of sampling a single atom at a time is that the conditional distribution posterior for an exponential prior on each atom and the normal likelihood is a truncated normal, enabling Gibbs sampling described in Section 4. This single random variable is called an "atom" and the set of all such atoms is referred to as the "atomic domain". The value of each matrix element of $\mathbf{A}$ is then given by

$$A_{i,k} = \sum_{l=1}^{\alpha^A_{i,k}} x^A_{i,k,l} \tag{2}$$

and similarly for $\mathbf{P}$. The atoms in the atomic domain are stored in ordered coordinates on a number line ($l^A_{i,k,l}$ for $\mathbf{A}$ and $l^P_{k,j,m}$ for $\mathbf{P}$), which is divided into bins that correspond to each matrix element

(Main Figure 1). The set of all atoms for one matrix is referred to as the "atomic domain". If the number of atoms is smaller than the number of matrix elements, this data structure reduces the memory required to keep track of each atom and provides an efficient structure to find all the atoms mapping to a single matrix elements. The prior distribution of atom coordinates is uniform, corresponding to an uniform prior for atom membership in each matrix element.

2 Update steps for the atomic prior

CoGAPS alternates between updating n_A atoms in the **A** and n_P atoms in the **P** matrices. The values of n_A and n_P are sampled from a Poisson distribution with parameter for the total number of atoms in the atomic domain for **A** (N_A) and in the atomic domain for **P** (N_P), respectively. Thus, on expectation all atoms in the domain are updated at each matrix-level iteration. The total number of such update steps is input as a parameter `nEquil` during the burn in stage (Section 5) and `nSamp` during the sampling stage.

In each of these n_A and n_P, we perform one of the four update steps to the respective atomic domains (Main Figure 1). We briefly describe these steps for **A** below, and note that they are defined similarly for **P**.

1. **Birth.** Create a single new atom in the atomic domain, so that $N_A \leftarrow N_A + 1$.
2. **Death or resize.** Change the value of a single atom $x^A_{i,k,l} \leftarrow x^A_{i,k,l} - \Delta x^A_{i,k,l}$, and removing it from the atomic domain so that $N_A \leftarrow N_A - 1$ if $x^A_{i,k,l} - \Delta x^A_{i,k,l} = 0$.
3. **Move.** Changing the location of a single atom ($x^A_{i,k,l}$) to a new location between adjacent atoms ($x^A_{m,n,p}$ and $x^A_{q,r,s}$) such that $l^A_{i,k,l} \in \left(l^A_{m,n,p}, l^A_{q,r,s}\right)$ on the atomic domain.
4. **Exchange.** Moving a portion of the value of a single atom ($x^A_{i,k,l}$) to another, adjacent atom ($x^A_{m,n,p}$) so that $x^A_{i,k,l} \leftarrow x^A_{i,k,l} + \Delta x$ and $x^A_{m,n,p} \leftarrow x^A_{i,k,l} - \Delta x$ where $\Delta x \in \left(-(x^A_{i,k,l} + x^A_{m,n,p}), x^A_{i,k,l} + x^A_{m,n,p}\right)$. Atoms may become small from exchange, but not exactly zero or removed from the atomic in order to maintain detailed balance.

At each of the n^A or n^P iterations, each of these four steps is chosen at random with 1/3 probability of either birth or death, 1/3 probability of move, and 1/3 probability of exchange. The relative probability of selecting birth or death is selected based on the Poisson prior. Recall for **A** that birth implies $N^A \leftarrow N^A + 1$, the sum of Poisson distributed random variables, and that under the Poisson distribution $P(N+1|N) = \frac{N}{(N+\lambda)}$ where λ is the Poisson parameter. Together, these three conditions suggest that $P\left(\text{birth}|N^A\right) = \frac{N^A}{(N^A+\alpha G k)}$ for the **A** atomic domain and that $P\left(\text{birth}|N^P\right) = \frac{N^P}{(N^P+\alpha k S)}$ for the **P** atomic domain. The probability of death or resize is then one minus the probability of birth. Metropolis Hastings sampling is used for the move step, whereas Gibbs sampling is used for the other three steps using the conditional distributions derived in Section 4.

3 Initialization

The atomic domains for both **A** and **P** are initialized without any atoms, so that $A_{i,j} = 0$ and $P_{k,l} = 0$. This limits the initial atomic update step to birth step, birth or death when there is at least one atom in the domain, and all four update steps when there are at least two atoms in the domain.

At these initial steps, the estimated fit to the data $\mu_{i,j} = A_{i,\cdot}P_{\cdot,j}$ will be zero for most values of i and j. Thus, these initial steps do not change the likelihood and are all accepted. This initialization effectively results in initial conditions which are a random sampling from the prior before Gibbs sampling.

4 Conditional distributions for Gibbs sampling

We would like to sample from Skilling's atomic domain using Gibbs sampling. We will assume that we are seeking the mass of an atom $x^A_{k,l}$ at $A_{k,l,j}$ for the **A** matrix and $x^P_{l,m,j}$ at $P_{l,m}$ for

the $\mathbf{P}$ matrix. We use the variable x in the derivations below to reduced the number of indices in the equations, as the associated matrix element can be clearly inferred from the context of each equation. The initial mass of this atom is x_0, which is 0 if we have decided to birth the atom and > 0 if we have decided to kill it. We retain this term so that we can derive the conditionals for birth and death in a single expression.

Determining the mass of x requires first computing the full conditional distribution $p\left(x|x_0, \mathbf{D}, \mathbf{A}, \mathbf{P}\right)$. To do this, we will first consider $P\left(\mathbf{A}, \mathbf{P}|\mathbf{D}\right)$ and examine the resulting distribution. We will begin by recalling that

$$p\left(\mathbf{A}, \mathbf{P}|\mathbf{D}\right) \propto p\left(\mathbf{D}|\mathbf{A}, \mathbf{P}\right) p\left(\mathbf{A}, \mathbf{P}\right). \tag{3}$$

Putting this in terms of an individual atom, we obtain

$$p\left(x|x_0, \mathbf{D}, \mathbf{A}, \mathbf{P}\right) \propto p\left(\mathbf{D}|x, x_0, \mathbf{A}, \mathbf{P}\right) p\left(x\right). \tag{4}$$

We assume that

$$p\left(\mathbf{D}|x, x_0, \mathbf{A}, \mathbf{P}\right) \sim N(\mathbf{M}, \mathbf{\Sigma}), \tag{5}$$

where $\mathbf{M}$ is the mock data matrix given by the product of $\mathbf{A}$ and $\mathbf{P}$ that incorporates the change in mass of the atom $x - x_0$ in the updated term. $\mathbf{\Sigma}$ is the covariance matrix for $\mathbf{D}$. The prior for the mass of each atom x is given by an exponential with parameters λ_A and λ_P, respectively.

In each case, the full conditional distribution simplifies to a normal distribution, which is truncated so that the value of the atom $x \geq 0$. Below follows the detailed derivation of this distribution for birth and resizing and exchange.

4.1 Conditional distribution for birth or resizing of atoms

4.1.1 Atomic domain for A

We will first explore the likelihood in more detail, assuming that the mass of the atom maps to $A_{k,l}$

$$p\left(\mathbf{D}|x, x_0, \mathbf{A}, \mathbf{P}\right) \propto \exp\left\{-\sum_i \sum_j \frac{1}{2\sigma_{i,j}^2}\left(D_{i,j} - \sum_p A_{i,p}P_{p,j} - \left(x - x_0\right)P_{l,j}\right)^2\right\}. \tag{6}$$

Since we are only concerned with computing the conditional for changes to $A_{k,l}$ we note that the other terms in $\mathbf{A}$ and $\mathbf{P}$ can be considered as parameters. As a result,

$$p\left(\mathbf{D}|x, x_0, \mathbf{A}, \mathbf{P}\right) \propto \exp\left\{-\sum_j \frac{1}{2\sigma_{k,j}^2}\left(D_{k,j} - \sum_p A_{k,p}P_{p,j} - \left(x - x_0\right)P_{l,j}\right)^2\right\}. \tag{7}$$

$$= \exp\left\{-\sum_j \frac{P_{l,j}}{2\sigma_{k,j}^2}\left(x - \left(\frac{D_{k,j} - \sum_p A_{k,p}P_{p,j} + x_0 P_{l,j}}{P_{l,j}}\right)\right)^2\right\}. \tag{8}$$

Let $\mu_{k,l,j}^A = \frac{D_{k,j} - \sum_p A_{k,p}P_{p,j} + x_0 P_{l,j}}{P_{l,j}}$ and $s_{k,l,j}^A = \frac{P_{l,j}^2}{2\sigma_{k,j}^2}$. Then, Equation (8) becomes

$$p\left(\mathbf{D}|x, x_0, \mathbf{A}, \mathbf{P}\right) \propto \exp\left\{-\sum_j s_{k,l,j}^A\left(x - \mu_{k,l,j}^A\right)^2\right\} \tag{9}$$

$$= \exp\left\{-\sum_j s_{k,l,j}^A\left(x^2 - 2\mu_{k,l,j}^A x + \mu_{k,l,j}^{A2}\right)\right\} \tag{10}$$

$$= \exp\left\{-\left(x^2\sum_j s_{k,l,j}^A - 2x\sum_j s_{k,l,j}^A\mu_{k,l,j}^A + \sum_j s_{k,l,j}^A\mu_{k,l,j}^{A2}\right)\right\} \tag{11}$$

$$\propto \exp\left\{-\sum_j s_{k,l,j}^A\left(x^2 - 2x\frac{\sum_j s_{k,l,j}^A\mu_{k,l,j}^A}{\sum_j s_{k,l,j}^A}\right)\right\}. \tag{12}$$

If we now incorporate the product with the exponential prior distribution for α,

$$p\left(x|x_0,\mathbf{D},\mathbf{A},\mathbf{P}\right) \propto \exp\left\{-\sum_j s^A_{k,l,j}\left(x^2-2x\frac{\sum_j s^A_{k,l,j}\mu^A_{k,l,j}}{\sum_j s^A_{k,l,j}}\right)\right\}\exp\left\{-\lambda_A x\right\} \tag{13}$$

$$= \exp\left\{-\sum_j s^A_{k,l,j}\left(x^2-x\left(2\frac{\sum_j s^A_{k,l,j}\mu^A_{k,l,j}}{\sum_j s^A_{k,l,j}}-\frac{\lambda_A}{\sum_j s^A_{k,l,j}}\right)\right)\right\} \tag{14}$$

$$\propto N\left(\frac{2\sum_j s^A_{k,l,j}\mu^A_{k,l,j}-\lambda_A}{2\sum_j s^A_{k,l,j}},\frac{1}{\sqrt{2\sum_j s^A_{k,l,j}}}\right). \tag{15}$$

Within the code, we store values of s and $s\times\mu$ used in Eq. 15 to avoid dividing by zero in cases where $P_{l,j}=0$.

4.1.2 Atomic domain for P

Here, we consider atoms whose mass maps to elements $P_{l,m}$. From the likelihood in Equation (6), we get

$$p\left(\mathbf{D}|x,x_0,\mathbf{A},\mathbf{P}\right) \propto \exp\left\{-\sum_i\frac{1}{2\sigma^2_{i,m}}\left(D_{i,m}-\sum_p A_{i,p}P_{p,m}-\left(x-x_0\right)A_{i,l}\right)^2\right\} \tag{16}$$

$$= \exp\left\{-\sum_i\frac{A_{i,l}}{2\sigma^2_{i,m}}\left(x-\left(\frac{D_{i,m}-\sum_p A_{i,p}P_{p,m}+x_0A_{i,l}}{A_{i,l}}\right)\right)^2\right\}. \tag{17}$$

If $\mu^P_{i,l,m}=\frac{D_{i,m}-\sum_p A_{i,p}P_{p,m}+x_0A_{i,l}}{A_{i,l}}$ and $s^P_{i,l,m}=\frac{A^2_{i,l}}{2\sigma^2_{i,m}}$,

$$p\left(\mathbf{D}|x,x_0,\mathbf{A},\mathbf{P}\right) \propto \exp\left\{-\sum_i s^P_{i,l,m}\left(x-\mu^P_{i,l,m}\right)^2\right\} \tag{18}$$

$$= \exp\left\{-\sum_i s^P_{i,l,m}\left(x^2-2\mu^P_{i,l,m}x+\mu^{P2}_{i,l,m}\right)\right\} \tag{19}$$

$$\propto \exp\left\{-\left(\sum_i s^P_{i,l,m}\right)\left(x^2-\frac{2\sum_i\mu^P_{i,l,m}s^P_{i,l,m}x}{\sum_i s^P_{i,l,m}}\right)\right\} \tag{20}$$

If we now incorporate the prior distribution for x

$$p\left(x|x_0,\mathbf{D},\mathbf{A},\mathbf{P}\right) \propto \exp\left\{-\left(\sum_i s^P_{i,l,m}\right)\left(x^2-\left(\frac{2\sum_i\mu^P_{i,l,m}s^P_{i,l,m}}{\sum_i s^P_{i,l,m}}\right)x\right)\right\}\exp\left\{-\lambda^P x\right\} \tag{21}$$

$$= \exp\left\{-\left(\sum_i s^P_{i,l,m}\right)\left(x^2-\left(\frac{2\sum_i\mu^P_{i,l,m}s^P_{i,l,m}-\lambda^P}{\sum_i s^P_{i,l,m}}\right)x\right)\right\} \tag{22}$$

$$\propto N\left(\frac{2\sum_i\mu^P_{i,l,m}s^P_{i,l,m}-\lambda^P}{2\sum_i s^P_{i,l,m}},\frac{1}{\sqrt{2\sum_i s^P_{i,l,m}}}\right) \tag{23}$$

4.2 Conditional distribution for exchange between neighboring atoms in the atomic domain

4.2.1 Exchange for A between $A_{k,l}$ and $A_{m,n}$ where $k\neq m$

We will refer to the atom corresponding to matrix element $A_{k,l}$ as x, the atom corresponding to the matrix element $A_{m,n}=x_{m,n}$, and x_0 and $x_{0,m,n}$ there initial values, respectively. The value of x after sampling is constrained such that $x\in(0,X)$ and $x_{m,n}=X-x$ where $X=x_0+x_{0,m,n}$.

If we consider the exponential prior, the exchange step will incorporate both matrix elements. That is,

$$\exp\left(-\lambda_A x\right)\exp\left(-\lambda_A(X-x)\right). \tag{24}$$

The x terms in this equation cancel, indicating that the conditional depends only on the likelihood. This occurs for all exchange steps, and thus is not described in the remaining subsections on this step.

From the likelihood in Equation (6), we get

$$p\left(\mathbf{D}|x, x_0, X, \mathbf{A}, \mathbf{P}\right) \propto \exp\left\{-\sum_j \frac{1}{2\sigma_{k,j}^2}\left(D_{k,j} - \sum_p A_{k,p}P_{p,j} - xP_{l,j}\right)^2\right\} \tag{25}$$

$$\times \exp\left\{-\sum_j \frac{1}{2\sigma_{m,j}^2}\left(D_{m,j} - \sum_p A_{m,p}P_{p,j} - (X-x)\,P_{n,j}\right)^2\right\}. \tag{26}$$

For simplicity of the equations, we consider only the terms inside of the exponential and formulate them as an equation for x to find the parameters of the truncated normal for value j in the summation.

$$\frac{\left[xP_{l,j} - \left(D_{k,j} - \sum_p A_{k,p}P_{p,j}\right)\right]^2}{2\sigma_{k,j}^2} + \frac{\left[xP_{n,j} - \left(XP_{n,j} + \sum_p A_{m,p}P_{p,j} - D_{m,j}\right)\right]^2}{2\sigma_{m,j}^2}. \tag{27}$$

Letting $\mu_{k,j} = D_{k,j} - \sum_p A_{k,p}P_{p,j}$ and $M_{m,n,j} = XP_{n,j} + \sum_p A_{m,p}P_{p,j} - D_{m,j}$, eq 27 simplifies to

$$\frac{\left(xP_{l,j} - \mu_{k,j}\right)^2}{2\sigma_{k,j}^2} + \frac{\left(xP_{n,j} - M_{m,n,j}\right)^2}{2\sigma_{m,j}^2}. \tag{28}$$

Combining terms, we can write this equation as

$$\frac{\left[\sigma_{m,j}^2P_{l,j}^2 + \sigma_{k,j}^2P_{n,j}^2\right]x^2 - 2\left[\sigma_{m,j}^2P_{l,j}\mu_{k,j} + \sigma_{k,j}^2P_{n,j}M_{m,n,j}\right]x}{2\sigma_{k,j}^2\sigma_{m,j}^2} \tag{29}$$

which can complete the square by

$$\frac{\left[\sigma_{m,j}^2P_{l,j}^2 + \sigma_{k,j}^2P_{n,j}^2\right]}{2\sigma_{m,j}^2\sigma_{n,j}^2}\left(x - \frac{\sigma_{m,j}^2P_{l,j}\mu_{k,j} + \sigma_{k,j}^2P_{n,j}M_{m,n,j}}{\sigma_{m,j}^2P_{l,j}^2 + \sigma_{k,j}^2P_{n,j}}\right)^2 \tag{30}$$

The parameters for the truncated normal can now follow the derivation used for the birth step described above.

4.2.2 Exchange for A between $A_{k,l}$ and $A_{k,n}$

Considering just the terms inside of the exponent for equation 26, in this case we will have instead

$$\sum_j \frac{\left(D_{k,j} - \sum_p A_{k,p}P_{p,j} - xP_{l,j} - (X-x)P_{n,j}\right)^2}{2\sigma_{k,j}^2} \tag{31}$$

Collecting the x terms and completing the square we get

$$\sum_j \frac{\left(P_{l,j} - P_{n,j}\right)^2}{2\sigma_{k,j}^2}\left[x - \frac{D_{k,j} - \sum_p A_{k,p}P_{p,j} - XP_{n,j}}{P_{l,j} - P_{n,j}}\right]^2 \tag{32}$$

we let $s_j = \frac{\left(P_{l,j} - P_{n,j}\right)^2}{2\sigma_{k,j}^2}$ and $\mu_j = \frac{D_{k,j} - \sum_p A_{k,p}P_{p,j} - XP_{n,j}}{P_{l,j} - P_{n,j}}$. The derivation for the terms of the truncated normal follow.

4.2.3 Exchange for P between $P_{k,l}$ and $P_{m,n}$ where $l \neq n$

The derivation for exchange steps in $\mathbf{P}$ follows that of the derivation for $\mathbf{A}$ in Section 4.2.1. In this case,

$$s_i = \frac{\sigma_{i,n}^2 A_{i,j}^2 + \sigma_{i,l}^2 A_{i,m}^2}{2\sigma_{i,l}^2 \sigma_{i,n}^2}, \tag{33}$$

$$\mu_i = \frac{\sigma_{i,n}^2 A_{i,k} \mu_{i,l} + \sigma_{i,l}^2 A_{i,m} M_{i,m,n}}{\sigma_{i,n}^2 A_{i,k}^2 + \sigma_{i,l}^2 A_{i,m}^2}, \tag{34}$$

where $\mu_{i,l} = D_{i,l} - \sum_p A_{i,p} P_{p,l}$ and $M_{i,m,n} = \sum_p A_{i,p} P_{p,n} + X A_{i,m} - D_{i,n}$.

4.2.4 Exchange for P between $P_{k,l}$ and $P_{m,l}$

The derivation for the exchange steps for **P** follows that of the derivation for **A** in Section 4.2.2. Thus, in this case

$$s_i = \frac{(A_{i,k} - A_{i,m})^2}{2\sigma_{i,l}^2}, \tag{35}$$

$$\mu_i = \frac{D_{i,l} - \sum_p A_{i,p} P_{p,l} - X A_{i,m}}{A_{i,k} - A_{i,m}}. \tag{36}$$

5 Annealing parameter

During the equilibration phase, we in fact wish to sample from the conditional distribution

$$p\left(x | x_0, \mathbf{D}, \mathbf{A}, \mathbf{P}\right) \propto p\left(\mathbf{D} | x, x_0, \mathbf{A}, \mathbf{P}\right)^{1/T} p\left(x\right), \tag{37}$$

where T is the annealing temperature. This has the effect of multiplying the term σ in each of the equations by a factor of T. As a result, the standard deviation s of the birth and resize terms are the only things to change by as follows.

$$s_{k,l,j}^A = \frac{P_{l,j}}{2T\sigma_{k,j}^2}, \text{ and} \tag{38}$$

$$s_{i,l,m}^P = \frac{A_{i,l}}{2T\sigma_{k,j}^2}. \tag{39}$$

A similar modification of the terms with $\sigma \leftarrow T\sigma$ will also occur in the exchange step, which will modify both the mean and standard deviation terms for this step.

References

[1] H Ishwaran and J S Rao. Spike and slab variable selection: Frequentist and bayesian strategies. *Annals of Applied Statistics*, 33:730–773, 2005.

[2] T D Moloshok, R R Klevecz, J D Grant, F J Manion, W F Speier, 4th, and M F Ochs. Application of bayesian decomposition for analysing microarray data. *Bioinformatics*, 18(4): 566–75, Apr 2002.

[3] S. Sibisi and J. Skilling. Prior distributions on measure space. *Journal of the Royal Statistical Society, B*, 59(1):217–235, 1997.