If Only My Posterior Were Normal: Introducing Fisher HMC

Adrian Seyboldt
adrian.seyboldt@gmail.com
PyMC-Labs

1. November 2024

Abstract

Hamiltonian Monte Carlo (HMC) has become a crucial tool in Bayesian inference, offering efficient exploration of complex, high-dimensional parameter spaces. However, HMC's performance is highly sensitive to the geometry of the posterior distribution, which is often poorly approximated by traditional mass matrix adaptations, especially in cases of non-normal or correlated posteriors. In this paper, we propose an adaptive framework for HMC that uses Fisher divergence to guide transformations of the parameter space, generalizing the concept of a mass matrix to arbitrary diffeomorphisms. By aligning the score function of the transformed posterior with those of a standard normal distribution, our method identifies transformations that adapt to the posterior's scale and shape. We develop theoretical foundations for these adaptive transformations, provide efficient implementation strategies, and demonstrate significant sampling improvements over conventional methods. Additionally, we introduce and evaluate nutpie, an implementation of our method for PyMC and Stan models and compare it with existing samplers across a suite of models. Our results show that nutpie delivers better sampling efficiency, particularly for challenging posteriors.

Keywords Bayesian Inference · Hamiltonian Monte Carlo · Mass Matrix Adaptation · Normalizing Flows · Fisher Divergence

1 Introduction

Hamiltonian Monte Carlo (HMC) is a powerful Markov Chain Monte Carlo (MCMC) method widely used in Bayesian inference for exploring complex posterior distributions. HMC can explore high-dimensional parameter spaces more efficiently than traditional MCMC techniques, making it popular in probabilistic programming libraries like Stan and PyMC. However, the performance of HMC depends critically on the parameterization of the posterior space, often framed in terms of a “mass matrix” that defines an inner product on this space. Properly tuning this parameterization is crucial for efficient sampling, but achieving optimal performance remains a challenging task, particularly for posterior distributions with complex geometries, such as those that exhibit strong correlations or funnel-like structures.

A common approach in HMC is to estimate a mass matrix based on the inverse of the posterior covariance, typically in a diagonalized form, to adjust for differences in scale across dimensions. While effective in many cases, this method has limitations when the posterior deviates significantly from normality. In cases where the posterior distribution has complex geometries, relying on a fixed mass matrix can lead to inefficient sampling, with trajectories that fail to capture the underlying structure of the parameter space. This inefficiency is especially pronounced when dealing with correlated parameters or posteriors that exhibit challenging topologies, such as those commonly encountered in hierarchical models and large-scale datasets.

To address these limitations, we propose an adaptive HMC framework that extends beyond the traditional concept of a mass matrix, using arbitrary diffeomorphisms to dynamically transform the parameter space. In this work, we utilize the Fisher divergence as a guiding criterion for these transformations, which allows us to adapt the geometry of the posterior space in a way that optimizes HMC's efficiency. By aligning the scores (derivatives of the log-density) of the transformed posterior with those of a standard normal distribution, we approximate an idealized parameterization that promotes efficient exploration.

Our approach is grounded in the observation that the Fisher divergence provides a meaningful metric to quantifying the discrepancy between the transformed posterior and a normal distribution. By minimizing this divergence, we can identify transformations that reduce the complexity of the posterior geometry. This adaptive framework generalizes the mass matrix concept, but also improves sampling efficiency if it is employed to find better mass matrices.

2 Adaptive Transformations in HMC: Motivation, Theory, and Examples

2.1 Motivation: Example with independent gaussian posterior

When we run HMC, we compute the derivatives of the posterior log density (the scores). They contain a lot of information about the posterior, but when we compute the mass matrix, we ignore this information.

To illustrate how useful the scores can be, let's start very simple, and assume our posterior is a one dimensional normal distribution $N(\mu, \sigma^2)$ with density $d\mu$, and let's assume we have two posterior draws x_1 and x_2, together with the covector (row-vector) of scores $\alpha_i = \frac{\partial}{\partial x_i} \log d\mu(x_i) = \sigma^{-2}(\mu - x_i)$. Based on this information we can directly compute μ and σ, and so identify the exact posterior, by solving for μ and σ. We get

$\mu = \bar{x} + \sigma^2\bar{\alpha}$ and $\sigma^2 = \mathrm{Var}[x_i]^{\frac{1}{2}}\mathrm{Var}[\alpha_i]^{-\frac{1}{2}}$, where $\bar{x}$ and $\bar{\alpha}$ are the sample means of x_i and α_i respectively. So we can compute the exact posterior and ideal mass matrix with no sample variance based on just two points!

This generalizes directly to multivariate normal posteriors $N(\mu, \Sigma)$. Let's assume we have $N+1$ linearly independent draws x_i, and the scores $\alpha_i = \Sigma^{-1}(x_i - \mu)$. The mean of these equations gives us $\mu = \bar{x} - \Sigma\bar{\alpha}$. It follows that $S = \Sigma^{-1}X$, where the i-th column of S is $\alpha_i - \bar{\alpha}$, and the i-th column of X is $x_i - \bar{x}$. And finally we get $SS^T = \mathrm{Cov}[\alpha_i] = \Sigma^{-1}XX^T\Sigma^{-1} = \Sigma^{-1}\,\mathrm{Cov}[x_i]\Sigma^{-1}$. We can recognize Σ as the geometric mean of the positive symmetric matrices $\mathrm{Cov}[x_i]$ and $\mathrm{Cov}[s_i]^{-1}$, so $\Sigma = \exp(\log(\mathrm{Cov}[x_i])/2 - \log(\mathrm{Cov}[\alpha_i]/2)$ or some expression with lots of matrix square roots (todo).

Given the scores, we can compute the parameters of a normal distribution exactly. Most posterior distributions are not multivariate normal distributions, or we would not have to run MCMC in the first place. It is quite common that they approximate normal distributions reasonably well, so this should indicate that the scores contain useful information we are currently neglecting.

2.2 Fisher divergence

todo: introduce notation $\frac{\nu}{\omega}$ for relative density.

todo: Write a second version of this without all the diffgeo...

Let (Y, g) be a Riemannian manifold with probability volume forms ω_1 and ω_2. They define scalar functions $p, q \in \Omega^0(Y)$ relative to the volume form of the metric tensor ω (so $\omega_1 = p\omega$ and $\omega_2 = q\omega$). We define the fisher divergence of ω_1 and ω_2 as

$$D_g(\omega_1, \omega_2) = \int \|d\log(p) - d\log(q)\|_g^2 d\omega_1. \tag{1}$$

Equivalently we can define $\omega_1 = z\omega_2$, and set

$$D_g(\omega_1, \omega_2) = \int \|d\log(z)\|_g^2 d\omega_1. \tag{2}$$

Note that D requires more structure on Y than the KL-divergence ($\int \log(z)d\omega_1$), as the norm depends on the metric tensor.

Given a second (non-Riemannian) manifold X with a probability volume form μ, and a diffeomorphism $F : Y \to X$, we can define the divergence between μ and ω_1 by pulling back μ to X, ie $D_g(F^*\mu, \omega_1)$.

The way we will apply this, is by noticing that the normal distributions are defined in terms of a metric tensor (or in $\mathbb{R}^n$ an inner product), because we can write the density as $\propto \exp(-\frac{1}{2}\mathrm{dist}(\mathrm{mean}, x)^2)$. So if we compute the fisher divergence between a normal distribution and some other distribution, we can always use the metric tensor of the normal distribution.

Notice, that if we were to compute the fisher divergence directly on X, we would have to take into account how the metric tensor changes with the transformation F:

$$D_g(F^*\mu, \omega_1) = D_{(F^{-1})^* g}\left(\mu, \left(F^{-1}\right)^* \omega_1\right) \tag{3}$$

2.3 Transformations of scores

In practice, we don't work with the measure μ directly, but with the density $\frac{\mu}{\lambda}$ (like the Radon-Nikodym derivative, also often written as $\frac{d\mu}{d\lambda}$) with respect to the Lebesgue measure λ, and we will have different Lebesgue measures on X and Y. So in order to compute $D_g(F^*\mu, \omega)$ we have to understand how to compute the score function $d\log\left(\frac{dF^*\mu}{d\lambda_N}\right)$ on Y if we have the score function $d\log\left(\frac{d\mu}{d\lambda_M}\right)$ on X.

$$\begin{aligned} d\log\left(\frac{F^*\mu}{\lambda_M}\right) &= d\log\left(\frac{F^*\mu}{F^*\lambda_N}\frac{F^*\lambda_N}{\lambda_M}\right) \\ &= F^* d\log\frac{\mu}{\lambda_N} + F^* d\log\left(\frac{\lambda_N}{\left(F^{-1}\right)^*\lambda_M}\right) \\ &= \hat{F}^*\left(d\log\frac{\mu}{\lambda_N}, 1\right) \end{aligned} \tag{4}$$

todo explain $\hat{F}$...

where $\log\left(\frac{\lambda_N}{\left(F^{-1}\right)^*\lambda_M}\right)$ is the log-determinant term in the change-of-variable formula.

We can simplify this a bit by introducing $\hat{F} : Y \to X \times \mathbb{R}$ where $\hat{F}(y) = \left(F(y), \log\left|\det\left(\frac{\partial F}{\partial y}\right)\right|\right)$

For instance in Jax:

```python
def F_and_logdet(y):
    """Compute the transformation F and its log determininant jacobian."""
    ...

def F_inv(x):
    """Compute the inverse of F."""
    ...

def F_pullback_score(x, s_x, logp):
    """Compute the transformed position, score and logp."""
    y = F_inv(x)
    (_, logdet), pullback_fn = jax.vjp(F_and_logdet, y)
    s_y = pullback_fn(s_x, 1.0)
    return y, s_y, logp + logdet
```

2.4 Transformed HMC

Given a posterior μ on X and a diffeomorphism $F : Y \to X$, we can run HMC on the transformed space Y:

```python
def hmc_proposal(rng, x, mu_density, F, step_size, n_leapfrog):
    logp_x, s_x = gradient_and_value(mu_density)(x)
    y, s_y, logp_y = F_pullback_score(x, s_x, logp_x)

    velocity = rng.normal(size=len(x))
    for i in range(n_leapfrog):
        # TODO just the usual leapfrog with identity mass matrix
        pass

    return
```

If F is an affine transformation, this simplifies to the usual mass-matrix based HMC. For instance, if we choose $F(x) = \sigma \odot x$, this corresponds to the mass matrix $\operatorname{diag}(\sigma^{-2})$.

HMC efficiency is famously dependent on the parametrization, so we know that this is much more or less efficient for some choices of F than for others. It is however not obvious what criterion we could use to decide how good a particular choice of F is. We need a loss function D that maps the diffeomorphism to a measure of difficulty for HMC.

This hard to quantify in general, but we can notice that the efficiency of HMC largely depends on the trajectory, and this trajectory does not depend on the density directly, but only the scores. We also know that HMC is efficient if the posterior is a standard normal distribution. So a reasonable loss function can ask how different the scores on the transformed space are from the scores of a standard normal distribution. If they match well, we will use the same trajectory we would use for a standard normal distribution ω, which we know to be a good trajectory. And because the standard normal distribution is defined in terms of an inner product, we already have a well-defined norm on the scores that we can use to evaluate their difference. But the fisher divergence $D(F^*\mu, \omega)$ measures exactly this difference. We have

$$\begin{aligned} D(F^*\mu, \omega) &= \int \left\| d \log \frac{F^*\mu}{\lambda_Y} - d \log \frac{\omega}{\lambda_Y} \right\|_g^2 F^*\mu \\ &= \int \left\| \frac{\partial}{\partial y} \log\left(\hat{F}^*\mu\right)(y) + y \right\|^2 dF^*\mu(y) \\ &= \int \left\| F^{\hat{*}} \frac{\partial}{\partial x} \log(d\mu(x)) + y \right\|^2 d(F^*\mu)(y) \end{aligned} \tag{5}$$

todo fix notation…

This is a special case of the fisher divergence between the transformed posterior and a standard normal distribution.

Given posterior draws x_i and corresponding scores α_i in the original posterior space X we can approximate this expectation as

$$\hat{D}_F = \frac{1}{N} \sum_i \left\| F^{\hat{*}} s_i + F^{-1}(x_i) \right\|^2 \tag{6}$$

Or in code:

```
def log_loss(F_pullback_scores, draw_data):
    draws, scores, logp_vals = draw_data
    pullback = vectorize(F_pullback_scores)
    draws_y, scores_y, _ = pullback(draws, scores, logp_vals)
    return log((draws_y + scores).sum(0).mean())
```

Note: Some previous literature (todo ref) proposed to minimize $\mathbb{E}[\alpha_x^T \alpha_x]$, which is similar, but does not solve the issue of choosing a well-defined inner product. But finding a good inner product is the whole point of mass matrix adaptation. If we pull pack the inner product of the standard normal distribution to X and use the corresponding inner product on the dual space of 1-forms, we end up with an equivalent definition for the loss function defined above.

2.5 Specific choices for the diffeomorphism F

For particular families of diffeomorphisms F, we can get more specific results.

2.5.1 Diagonal mass matrix

If we choose $F_{\sigma,\mu} : Y \to X$ as $x \mapsto y \odot \sigma + \mu$, we get the same effect as diagonal mass matrix estimation.

In this case, the fisher divergence reduces to

$$\hat{D}_{\sigma,\mu} = \frac{1}{N} \sum_i \left\| \sigma \odot \alpha_i + \sigma^{-1} \odot (x_i - \mu) \right\|^2 \tag{7}$$

This is a special case of the affine transformation in Section 6.1 and minimal if $\sigma^2 = \text{Var}[x_i]^{\frac{1}{2}} \text{Var}[\alpha_i]^{-\frac{1}{2}}$ and $\mu = \bar{x}_i + \sigma^2 \bar{s}_i$. So we recover the same result we got in Section 2.1. It is very easy to compute (also using an online algorithm for the variance to avoid having to store the x_i and α_i values), and therefore the default in nutpie. In a Section 5 we will show some benchmarks to compare its performance.

Some theoretical results for normal posteriors

If the posterior is $N(\mu, \Sigma)$, then the minimizers $\hat{\mu}$ and $\hat{\sigma}$ of $\hat{D}_F$) converge to μ and $\exp(\frac{1}{2} \log \text{diag}(\Sigma) - \frac{1}{2} \log \text{diag}(\Sigma^{-1})$ respectively. This is a direct consequence of $\text{Cov}[x_i] \to \Sigma$ and $\text{Cov}[\alpha_i] \to \Sigma^{-1}$.

D_F converges to $\sum_i \lambda_i + \lambda_i^{-1}$, where λ_i are the generalized eigenvalues of Σ with respect to $\text{diag}(\hat{\sigma}^2)$, so large and small eigenvalues are penalized. When we choose $\text{diag}(\Sigma)$ as mass matrix, we effectively minimize $\sum_i \lambda_i$, and only penalize large eigenvalues. If we choose $\text{diag}(\mathbb{E}(\alpha\alpha^T))$ we effectively minimize $\sum \lambda_i^{-1}$ and only penalize small eigenvalues. But based on some theoretical work for multivariate normal distributions, we know that both large and small eigenvalues make HMC less efficient. (todo ref)

We can use the result in (todo ref) to evaluate the different diagonal mass matrix choices on various gaussian posteriors, with different numbers of observations. Figure todo shows the resulting condition numbers of the posterior as seen by the sampler in the transformed space.

2.5.2 Full mass matrix

We choose $F_{A,\mu}(y) = Ay + \mu$. This corresponds to a mass matrix $M = \left(AA^T\right)^{-1}$. Because as we will see $\hat{D}_F$ only depends on AA^T and μ, we can restrict A to be symmetric positive definite.

We get

$$\hat{D}[F] = \frac{1}{N} \sum \left\| A^T s_i + A^{-1}(x_i - \mu) \right\|^2 \tag{8}$$

which is minimal if $AA^T \operatorname{Cov}[x_i] AA^T = \operatorname{Cov}[\alpha_i]$ (for a proof, see Section 6.1), and as such corresponds again to our earlier derivation in Section 2.1. If the two covariance matrices are full rank, we get a unique minimum at the geometric mean of $\operatorname{Cov}[x_i]$ and $\operatorname{Cov}[s_i]$.

2.5.3 Diagonal plus low-rank

If the number of dimensions is larger than the number of draws, we can add regularization terms. And to avoid $O(n^3)$ computation costs, we can project draws and scores into the span of x_i and α_i, compute the regularized mean in this subspace and use the mass matrix .... If we only store the components, we can avoid $O(n^2)$ storage, and still do all operations we need for HMC quickly. To further reduce computational cost, we can ignore eigenvalues that are close to one.

todo

Implemented in nutpie with `nutpie.sample(model, low_rank_modified_mass_matrix=True)`

2.5.4 Model specific diffeomorphisms

Sometimes we can suggest useful families of diffeomorphism based on the model itself. A common reparametrization for instance is the non-centered parametrization, where we change a model from the centered parametrization

$$x \sim N(0, \sigma^2) \tag{9}$$

into the non-centered parametrization

$$\begin{aligned} z &\sim N(0, 1) \\ x &= z \odot \sigma \end{aligned} \tag{10}$$

We can generalize this for any choice of $0 \leq t \leq 1$ to (todo add ref, ask Aki)

$$\begin{aligned} z &\sim N(0, \sigma^{2t}) \\ x &= z \odot \sigma^{1-t} \end{aligned} \tag{11}$$

This suggests $F_{t(\sigma,z)} = (\sigma, \sigma^{1-t} z)$...

2.5.5 Normalizing flows

Normalizing flows provide a large family of diffeomorphisms that we can use to transform our posterior. We have rather strong requirements for the flows however: We need forward and inverse transformations, and we need to be able to compute the log determinant of the jacobian of the transformation efficiently. A well studied familiy of normalizing flows that provides all of those is RealMVP (todo ref).

For our experiments we used the library flowjax that implemnets RealMVP and other normalizing flows in jax. I slightly changed the adaptation schema from the usual nutpie algorithm (described in more detail in the next section), so that for the first couple of windows only diagonal mass matrix adaptation is used, and only after 150 draws do we start to fit a normalizing flow.

We then repeatedly run an adam optimizer on a window of draws, and use the updated normalizing flow to sample.

Especially for larger models the size of our training data set seems to be very important, so I included the full trajectory of the HMC sampler as training data, not just the draws themselves. I think this might be usful to do even if the amount of training data is not a limiting factor, as we would like that the transformed posterior matches the normal distribution everywhere the HMC sampler evaluaties it, not just at those points it accepts as draws.

So far I have not tested this systematically, but my experiments suggest that this can help us to sample a wide range of posterior distributions a lot more efficiently, or also to sample many distributions that were previously not possible to sample without extensive manual reparametrizations. It also comes with a large computational cost however, as the optimization itself can take a long time. Luckily, it seems to run relatiely efficiently on GPUs, so that even if the logp function itself does not lend itself to evaluation on GPUs, we can still spend most of the comutational time running scalable code. Often, the number of gradient evaluations that are necessary to sample even complicated posterior distributions decrease a lot.

With the current implementation I did not have too much luck with problems however, that have significantly more than around 1000 parameters. I am running into memory issues (I think those are probably an implementation issue and not a fundamental issue with the algorithm however) and the effectiveness of the normalizing flows, or our ability to fit them properly with the limited number of evaluation points seems to decrease.

I hope that this issue can be overcome by choosing normalizing flow layers in a way that takes model structure into account, or by adding layers one after the other, based on which parameters seem to not fit the normal distirbuito well yet.

Current code: https://github.com/pymc-devs/nutpie/pull/154

3 Adaptation schema to learn the diffeomorphism

Whether we adapt a mass matrix using the posterior variance as Stan does, or if we use any of the bijections based on the fisher divergence defined above, we always have the same problem: In order to generate posterior draws we need the mass matrix (or the bijection), but to estimate the mass-matrix/bijection we need posterior draws.

There is a well known way out of this conundrum: We start sampling with same initial transformation, and collect a number of draws. Based on those draws, we estimate a better transformation, and repeat. This adaptation-window approach has long been used in the major implementations of HMC, and remained largely unchanged for a number of years. PyMC, numpyro and blackjax all mostly use the same details as Stan, with at most minor modifications.

There are howerver I think a couple of minor changes that seem to improve the effiencey of this schma significantly. Except for the last section, where I discuss adaptation schemas for normalizing flows, I will assume that we adapt a mass matrix (or equivalently an affine bijection).

3.1 Choice of initial position

Stan draws initial points independently for each chain from the interval $(-1, 1)$ on the unconstrained space. I don't have any clear data to suggest so, but for some time PyMC has initialized using draws from the prior instead, and it seems to me that this tends to be more robust. This is of course only possible for model definition languages that allow prior sampling, and would for instance be difficult to implement in Stan.

3.2 Choice of initial diffeomorphism

Stan starts the first adaptation with an identity mass matrix. The very first HMC trajectories seem to be overall much more reasonable if we use $M = \operatorname{diag}(\alpha_0^T \alpha_0)$ instead. This also makes the initialization independent of variable scaling.

3.3 Accelerated window based adaptation

Stan and other sampler do not run vanilla HMC, but often NUTS, where the length of the trajectory is choosen automatically. This can make it very costly to generate draws if the mass matrix is not adapted well, because in those cases we often use a large number of HMC steps for each draw (in the 100s or typically up to 1000). Very early on during sampling we have a large intcentive to use available information obout the posterior as quickly as possible, to avoid these long trajectories. By default Stan starts adaptation with a step-size adaptation window, (50 draws todo check), where we do not change the mass matrix at all. It is then followed by a mass matrix adaptation window (100 draws? todo), where we generate draws for the next mass matrix update, but still use the initial mass matrix for sampling.

It is not uncommon that trajectories are very long during these first 150(todo) draws, and drop significantly after the first update. And because the trajectory lengths can easily vary by a factor of 10 or 100 between these phases, the draws before the first mass matrix change can take a sizable percentage of the total sampling time.

```python
class MassMatrixEstimator:
    def update(self, position, score):
        ...
    def current(self) -> MassMatrix:
        ...
    def num_points(self) -> int:
        ...

class StepSizeAdapt:
    def reset(self):
        ...
    def update(self, accept_statistic):
        ...
    def current_warmup(self) -> float:
        ...
    def current_best(self) -> float:
        ...
```

```python
def warmup(num_warmup, num_early, num_late, early_switch_freq, late_switch_freq):
    position, score = draw_from_prior()
    foreground_window = MassMatrixEstimator()
    foreground_window.update(position, score)
    background_window = MassMatrixEstimator()
    step_size_estimator = StepSizeAdapt(position, score)
    first_mass_matrix = True

    for draw in range(num_warmup):
        is_early = draw < num_early
        is_late = num_warmup - draw < num_late

        mass_matrix = foreground_window.current()
        step_size = step_size_estimator.current_warmup()
        (
          accept_stat, accept_stat_sym, position, score,
          diverging, steps_from_init
        ) = hmc_step(mass_matrix, step_size, position, score)

        # Early on we ignore diverging draws that did not move
        # several steps. They probably just used a terrible step size
        ok = (not is_early) or (not diverging) or (steps_from_init > 4)
        if ok:
            foreground_window.update(position, score)
            background_window.update(position, score)

        if is_late:
            step_size_estimator.update(accept_stat_sym)
            continue

        step_size_estimator.update(accept_stat)

        switch_freq = early_switch_freq if is_early else late_switch_freq
        remaining = num_warmup - draw - num_late
        if (remaining > late_switch_freq
            and background_window.num_points() > switch_freq
        ):
            foreground_window = background_window
            background_window = MassMatrixEstimator()
            if first_mass_matrix:
                step_size_estimator.reset()
            first_mass_matrix = False
```

Constant step size adaptation with dual averaging. Overlapping windows, so that we can switch to a better estimate quickly.

Intuition: We want to update quickly, and not use an old mass matrix estimate at a point when we have more information and could already compute a better estimate. We could just use a tailing window and update in each step with the previous k draws. This is computationally inefficient (unless the logp function is very expensive), and can not easily be implemented as a streaming estimator (see below for more details). But if we use several overlapping estimation windows, we can compromise between optimal information usage and computational cost, and still use streaming mass matrix estimators.

Three phases:

- Initial phase with small window size to find the typical set. Discard duplicate draws.
- Main phase with longer windows
- Final phase with constant mass matrix: Only step size is adapted. Use a symmetric estimate of the acceptance statistic (todo ref stan discourse).

This scheme can be used with arbitrary mass matrix estimators. If the estimator allows a streaming (ref) implementation, we do not need to store the draws within each window.

4 Implementation in nutpie

todo

5 Experimental evaluation of nutpie

We run nutpie and cmdstan on posteriordb to compare performance in terms of effective sample size per gradient evaluation and in terms of effective sample size per time…

Code is here:

todo

6 Appendix

6.1 Minimize Fisher divergence for affine transformations

$D[F]$ for $F(y) = Ay + \mu$ is minimal if $\Sigma \operatorname{Cov}[\alpha]\Sigma = \operatorname{Cov}[x]$ and $\mu = \bar{x} + \Sigma\bar{\alpha}$, where $\Sigma = AA^T$:

We collect all α_i in the columns of G, and all x_i in the columns of X. Let e we the vector containing only ones. And let $\Sigma = AA^T$.

$$D = \frac{1}{N}\left\|A^T G + A^{-1}(X - \mu e^T)\right\|_F^2 \tag{12}$$

6.1.1 Find $\hat{\mu}$

Take the differential with respect to only μ:

$$\begin{aligned} dD &= -\frac{2}{N}\operatorname{tr}\left[\left(A^T G + A^{-1}(X - \mu e^T)\right)^T A^{-1} d\mu e^T\right] \\ &= -\frac{2}{N}\operatorname{tr}\left[e^T\left(A^T G + A^{-1}(X - \mu e^T)\right)^T A^{-1} d\mu\right] \end{aligned} \tag{13}$$

At a minimum of D this has to be zero for all $d\mu$, which is the case iff

$$e^T\left(A^T G + A^{-1}(X - \mu e^T)\right)^T A^{-1} = 0 \tag{14}$$

It follows that

$$\mu = \bar{x} + \Sigma\bar{\alpha} \tag{15}$$

6.1.2 Find $\hat{A}$

Take the differential of D with respect to A

$$dD = \frac{2}{N}\operatorname{tr}\Big[\big(A^TG + A^{-1}(X-\mu e^T)\big)^T\big(dA^TG - A^{-1}dAA^{-1}(X-\mu e^T)\big)\Big] \tag{16}$$

Using the result for μ we get $X - \mu e^T = \tilde{X} - \Sigma\bar{\alpha}e^T$. This, and using cyclic and transpose properties of the trace gives us

$$\begin{aligned} dD = {} & \frac{2}{N}\operatorname{tr}\big[\big(A^TG + A^{-1}\big(\tilde{X} - \Sigma\bar{\alpha}e^T\big)\big)G^TdA\big] \\ & + \frac{2}{N}\operatorname{tr}\Big[A^{-1}\big(\Sigma\bar{\alpha}e^T - \tilde{X}\big)\big(A^TG + A^{-1}\big(\tilde{X} - \Sigma\bar{\alpha}e^T\big)\big)^TA^{-1}dA\Big] \end{aligned} \tag{17}$$

This is zero for all dA iff

$$\begin{aligned} 0 &= \big(A^TG + A^{-1}\big(\tilde{X} - \Sigma\bar{\alpha}e^T\big)\big)G^T + A^{-1}\big(\Sigma\bar{\alpha}e^T - \tilde{X}\big)\big(A^TG + A^{-1}\big(\tilde{X} - \Sigma\bar{\alpha}e^T\big)\big)^TA^{-1} \\ &= A^T\tilde{G}G^T + A^{-1}\tilde{X}G^T + \big(A^T\bar{\alpha}e^T - A^{-1}\tilde{X}\big)\big(\tilde{G}^T + \tilde{X}^T\Sigma^{-1}\big), \end{aligned} \tag{18}$$

where $\tilde{X} = X - \bar{x}e^T$, the matrix with centered x_i in the columns, and $\tilde{G} = G - \bar{\alpha}e^T$. Because $e^T\tilde{X}^T = e^T\tilde{G}^T = 0$ we get

$$A^T\tilde{G}G^T + A^{-1}\tilde{X}G^T - A^{-1}\tilde{X}\tilde{G}^T - A^{-1}\tilde{X}\tilde{X}^T\Sigma^{-1} = 0 \tag{19}$$

Or

$$\begin{aligned} & \Big(\tilde{G} + \Sigma^{-1}\tilde{X}\Big)G^T - \Sigma^{-1}\tilde{X}\tilde{G}^T - \Sigma^{-1}\tilde{X}\tilde{X}^T\Sigma^{-1} \\ & = \tilde{G}G^T + \Sigma^{-1}\tilde{X}e\bar{\alpha}^T - \Sigma^{-1}\tilde{X}\tilde{X}^T\Sigma^{-1} \\ & = \tilde{G}\tilde{G}^T - \Sigma^{-1}\tilde{X}\tilde{X}^T\Sigma^{-1} \end{aligned} \tag{20}$$