The Third Moment Tensor Method With Principal Components and Basis Expansion

W.D. Brinda* Ruchira Ray†

Abstract

In recent years, researchers have developed the tensor method to approximate closed-form solutions for method of moments estimation procedures involving third-order moments. The tensor method indicates a way of performing estimation for problems in which maximum-likelihood is computationally intractable due to a highly multi-modal likelihood surface. However, the tensor method has complications and instabilities that limit its use in practice. One notable issue is the requirement that the true parameters must be linearly independent, implying that the dimension must be at least as large as the number of parameters. Here, we present two variants of the tensor method that engage with this issue. First, we describe an approach using principal components which are not linearly independent but make the ingredients of the tensor method more familiar and intuitive. Second, we describe the process of embedding the vectors in a higher-dimensional space to apply a tensor method to the embedded vectors. Finally, we demonstrate our new approaches in the context of estimating Gaussian mixtures. The principal components version of the algorithm results in a simpler estimation of the component parameters' third-order moments. The basis expansion example demonstrates how to devise additional variables and estimate the relevant moments in the higher-dimensional space.

Key Words: tensor method, Gaussian mixtures, basis expansion, method of moments, principal components

1. Introduction

Method of moments estimation procedures choose a model distribution whose moments match empirical moments of the data. For any distribution on $\mathbb{R}^d$, the first moments comprise a vector in $\mathbb{R}^d$. A particular value of the first moment may uniquely correspond to a model distribution if the model has fewer than d parameters. The second moments comprise a positive semi-definite matrix in $\mathbb{R}^{d\times d}$. Again, a particular combination of first and second moments may correspond to a unique model distribution if the model has few enough parameters. First and second moments are not sufficient to identify distributions within higher-dimensional models, but one can then make reference to higher moments. Techniques have recently been developed to relate certain models' parameters to the generating distribution's tensor of third moments and to efficiently find a model distribution corresponding approximately to a given set of first, second, and third moments.

The idea at the heart of the new tensor methods comes from [Chang, 1996] in the context of Markov models; the idea's generality and broader usefulness were revealed in [Anandkumar et al., 2012] and [Anandkumar et al., 2014]. Essentially, the "tensor trick" involves transforming a third-order tensor such that it becomes a sum of rank-one tensors built from orthonormal vectors that relate meaningfully to vectors of interest, such as model parameters. Section 2 first presents the algorithm in the abstract context of discovering a set of fixed vectors, adapted from explanations in [Anandkumar et al., 2012] and Section 2 of [Hsu and Kakade, 2013]. We then describe a variant of the algorithm that starts by centering the vectors which we argue is more intuitive and which makes the third-order tensors

*Quantitations LLC

†Department of Statistics, Columbia University, 1255 Amsterdam Avenue, New York, NY 10027

more convenient to estimate. We also describe a variant of the algorithm that embeds the vectors in a higher-dimensional space; this makes it possible to discover more component vectors than there are dimensions in the original data. In Section 3, we use our variants of the tensor method to estimate the component means of a Gaussian mixture. With the first variant, we estimate a Gaussian mixture whose components have an unknown variance. With the second variant, we estimate a five-component Gaussian mixture in $\mathbb{R}^2$. All code used in our examples is available at `www.quantitations.com/research`.

2. Tensor Method for Finding Vectors

We start with an abstract formulation of what the tensor method accomplishes. Let $\{\mu_1, \ldots, \mu_k\}$ be linearly independent vectors in $\mathbb{R}^d$, and let $\mathbb{P}$ be the discrete distribution on these points with all probabilities equal to $1/k$. Using empirical moments of this set of points, we will see how to calculate the original vectors.

2.1 Using Second and Third-order Moments

If the second-order moments

$$S_{\mathbb{P}} := \tfrac{1}{k} \sum_j \mu_j \otimes \mu_j$$

and the third-order moments

$$T_{\mathbb{P}} := \tfrac{1}{k} \sum_j \mu_j \otimes \mu_j \otimes \mu_j$$

are both known, then the vectors $\{\mu_1, \ldots, \mu_k\}$ can be calculated as follows.

Noticing that $S_{\mathbb{P}}$ has rank k, let $\sum_{j=1}^{k} \lambda_j q_j \otimes q_j$ comprise a decomposition of $S_{\mathbb{P}}$ with $q_1, \ldots, q_k$ orthonormal and $\lambda_1 \geq \ldots \geq \lambda_k > 0$. The square root of the Moore-Penrose inverse of $S_{\mathbb{P}}$ is

$$S_{\mathbb{P}}^{-1/2} := \sum_{j=1}^{k} \frac{1}{\sqrt{\lambda_j}} q_j \otimes q_j.$$

The so-called "whitened" vectors $\{\frac{1}{\sqrt{k}} S_{\mathbb{P}}^{-1/2} \mu_1, \ldots, \frac{1}{\sqrt{k}} S_{\mathbb{P}}^{-1/2} \mu_k\}$ are orthonormal because the sum of their outer products is an orthogonal projection operator (Lemma 1)

$$\begin{aligned}
\sum_j \left(\tfrac{1}{\sqrt{k}} S_{\mathbb{P}}^{-1/2} \mu_j\right) \otimes \left(\tfrac{1}{\sqrt{k}} S_{\mathbb{P}}^{-1/2} \mu_j\right) &= S_{\mathbb{P}}^{-1/2} \left[\tfrac{1}{k} \sum_j \mu_j \otimes \mu_j\right] S_{\mathbb{P}}^{-1/2} \\
&= S_{\mathbb{P}}^{-1/2} \left[\sum_j \lambda_j q_j \otimes q_j\right] S_{\mathbb{P}}^{-1/2} \\
&= \sum_j \lambda_j (S_{\mathbb{P}}^{-1/2} q_j) \otimes (S_{\mathbb{P}}^{-1/2} q_j) \\
&= \sum_j \lambda_j \left(\tfrac{1}{\sqrt{\lambda_j}} q_j\right) \otimes \left(\tfrac{1}{\sqrt{\lambda_j}} q_j\right) \\
&= \sum_j q_j \otimes q_j.
\end{aligned}$$

Next consider the transformation formed by applying a unit vector v into $T_{\mathbb{P}}$ then "whitening" both sides and rescaling:

$$S_{\mathbb{P}}^{-1/2}[T_{\mathbb{P}} \cdot v]S_{\mathbb{P}}^{-1/2} = S_{\mathbb{P}}^{-1/2}\left[\tfrac{1}{k}\sum_j \langle v, \mu_j\rangle \mu_j \otimes \mu_j\right] S_{\mathbb{P}}^{-1/2}$$
$$= \sum_j \underbrace{\langle v, \mu_j\rangle}_{\text{“}\gamma_j\text{”}} \left(\tfrac{1}{\sqrt{k}} S_{\mathbb{P}}^{-1/2}\mu_j\right) \otimes \left(\tfrac{1}{\sqrt{k}} S_{\mathbb{P}}^{-1/2}\mu_j\right). \quad (1)$$

Because $\{\frac{1}{\sqrt{k}} S_{\mathbb{P}}^{-1/2}\mu_1, \ldots, \frac{1}{\sqrt{k}} S_{\mathbb{P}}^{-1/2}\mu_k\}$ are orthonormal, this represents a spectral decomposition. Therefore, as long as the eigenvalues are distinct, a set of k eigenvectors $\{w_1, \ldots, w_k\}$ of $S_{\mathbb{P}}^{-1/2}[T_{\mathbb{P}} \cdot v]S_{\mathbb{P}}^{-1/2}$ are equal to the whitened vectors $\{\frac{1}{\sqrt{k}} S_{\mathbb{P}}^{-1/2}\mu_1, \ldots, \frac{1}{\sqrt{k}} S_{\mathbb{P}}^{-1/2}\mu_k\}$ except for possibly different signs. Transforming w_j by $\sqrt{k}S_{\mathbb{P}}^{1/2} = \sqrt{k}\sum_j \sqrt{\lambda_j} q_j \otimes q_j$ undoes the whitening and gives us either μ_j or $-\mu_j$. If γ_j (see Equation 1) equals the inner product of v with the proposed mean $\sqrt{k}S_{\mathbb{P}}^{1/2}w_j$, then the sign of u_j does not need to be reversed. Otherwise γ_j is $-\langle v, \sqrt{k}S_{\mathbb{P}}^{1/2}w_j\rangle$, and we conclude that $\mu_j = -\sqrt{k}S_{\mathbb{P}}^{1/2}w_j$. It suffices to compare the sign of $\langle v, S_{\mathbb{P}}^{1/2}w_j\rangle$ with that of the corresponding eigenvalue γ_j. This can be expressed conveniently in terms of an intermediate vector u_j where

$$u_j := \sqrt{k}S_{\mathbb{P}}^{1/2}w_j$$
$$\mu_j = \text{sign}(\gamma_j\langle v, u_j\rangle)\, u_j.$$

The unit vector v can be generated uniformly at random in order to ensure that the eigenvalues are distinct with probability 1.

We note that in some cases it preferable to *directly estimate* $T_{\mathbb{P}} \cdot v$ rather than estimating the third-order tensor then applying v into that estimate. We will make use of this approach in Section 3.2.

2.2 Using Central Moments

Additional intuition is provided by a variant of this algorithm that uses $\bar{\mu} := \frac{1}{k}\sum_j \mu_j$ along with the second and third-order *central* moments

$$\Sigma_{\mathbb{P}} := \tfrac{1}{k}\sum_j (\mu_j - \bar{\mu}) \otimes (\mu_j - \bar{\mu})$$

and

$$\Gamma_{\mathbb{P}} := \tfrac{1}{k}\sum_j (\mu_j - \bar{\mu}) \otimes (\mu_j - \bar{\mu}) \otimes (\mu_j - \bar{\mu}).$$

A spectral decomposition of $\Sigma_{\mathbb{P}}$ as $\sum_{j=1}^k \lambda_j q_j \otimes q_j$ is a familiar step in principal components analysis. The eigenvectors provide the directions that spread the points out from most to least, and the corresponding eigenvalues tell us the variances along those directions. Centering can also simplify the task of estimating moments as exemplified in the upcoming discussion of Gaussian mixtures.

A complication with this approach is the fact that the centered points cannot be linearly independent, as they sum to zero. To circumvent this issue, assuming eigenvalues $\lambda_1, \ldots, \lambda_{k-1}$ are positive, imagine adding the eigenvector q_k to each centered point. The

resulting second moments $\frac{1}{k}\sum_j(\mu_j - \bar{\mu} + q_k) \otimes (\mu_j - \bar{\mu} + q_k)$ simplify to $\Sigma_{\mathbb{P}} + q_k \otimes q_k$ which shares the same eigenvectors as $\Sigma_{\mathbb{P}}$ except that its eigenvalue corresponding to q_k is 1 rather than zero. Using q_1 as the vector that we apply into the third-order tensor of q_k-shifted centered vectors, we'll ultimately need the eigenvectors of

$$[\Sigma_{\mathbb{P}}^{-1/2} + q_k \otimes q_k] \left[\frac{1}{k} \sum_j (\mu_j - \bar{\mu} + q_k) \otimes (\mu_j - \bar{\mu} + q_k) \otimes (\mu_j - \bar{\mu} + q_k) \cdot q_1 \right] [\Sigma_{\mathbb{P}}^{-1/2} + q_k \otimes q_k]$$

which can be shown to simplify to

$$\Sigma_{\mathbb{P}}^{-1/2}[\Gamma_{\mathbb{P}} \cdot q_1]\Sigma_{\mathbb{P}}^{-1/2} + \sqrt{\lambda_1}[q_1 \otimes q_k + q_k \otimes q_1].$$

If it has k distinct positive eigenvalues $\gamma_1, \ldots, \gamma_k$, then its eigenvectors $w_1, \ldots, w_k$ provide the solution

$$\begin{aligned} u_j &:= \sqrt{k} \left[\Sigma_{\mathbb{P}}^{1/2} + q_k \otimes q_k \right] w_j \\ \mu_j - \bar{\mu} + q_j &= \text{sign}(\gamma_j \langle q_1, u_j \rangle)\, u_j. \end{aligned}$$

Another choice of eigenvector $q_2, \ldots, q_{k-1}$ can be used instead if needed for distinct eigenvalues.

The method can be expressed a bit more simply by recalling that every $\mu_j - \bar{\mu}$ is orthogonal to q_k. As a result, there is no need to keep track of the coefficient with respect to q_k, so the solution is also represented

$$\begin{aligned} \tilde{u}_j &:= \sqrt{k}\Sigma_{\mathbb{P}}^{1/2} w_j \\ \mu_j - \bar{\mu} &= \text{sign}(\gamma_j \langle q_1, \tilde{u}_j \rangle)\, \tilde{u}_j. \end{aligned}$$

2.3 Using Basis Expansion

If the number of points is smaller than the dimension of the space, the points can still be recovered from the moments of their embeddings into a higher-dimensional space. If the mapping f from $\mathbb{R}^d$ to a vector space results in linearly independent images $f(\mu_1), \ldots, f(\mu_k)$, then the image vectors $f(\mu_1), \ldots, f(\mu_k)$ can be calculated from their moments. With suitable choices of f, one can determine $\mu_1, \ldots, \mu_k$ from their images. In applications, care must be taken to design a mapping for which the moments of the higher-dimensional images can be successfully estimated, as we will see in Section 3.2.

3. Third Moment Tensor Method for Estimating Model Parameters

The tensor method can be used when the second and third-order moments of a set of parameter vectors can be estimated from the data, in which case those estimates can be substituted into the algorithm.

3.1 Gaussian Mixture with Unknown Variance

As an example, let P be an equally-weighted mixture of k Gaussian component distributions each having covariance $\sigma^2 I$ for some unknown common σ^2. Let $\{\mu_1, \ldots, \mu_k\}$ denote the components' unknown locations.

It is straightforward to relate moments of P to moments of $\mathbb{P}$, the discrete distribution defined by $\{\mu_1, \ldots, \mu_k\}$. Let J be uniform on $\{1, \ldots, k\}$ and let $Z \sim N(0, \sigma^2 I)$ be independent of J in order to represent X as $\mu_J + Z$. By iterated expectation, the expected

value of a draw from P is exactly equal to the average of the component means: $\mathbb{E}_{X\sim P}X = \bar{\mu}$. Furthermore, by the law of total covariance that results from conditioning on J, $\Sigma_P = \Sigma_{\mathbb{P}} + \sigma^2 I$. Notice the implication that P and $\mathbb{P}$ share the exact same principal component vectors and that their variances differ by exactly σ^2. For the third-order central moments,

$$\begin{aligned}\Gamma_P &:= \mathbb{E}(X-\bar{\mu})\otimes(X-\bar{\mu})\otimes(X-\bar{\mu}) \\ &= \mathbb{E}(\mu_J-\bar{\mu}+Z)\otimes(\mu_J-\bar{\mu}+Z)\otimes(\mu_J-\bar{\mu}+Z) \\ &= \underbrace{\mathbb{E}(\mu_J-\bar{\mu})\otimes(\mu_J-\bar{\mu})\otimes(\mu_J-\bar{\mu})}_{\Gamma_{\mathbb{P}}}\end{aligned}$$

because each of the other terms has expectation zero. When using ordinary third-order moments rather than central moments, the relationship is more complicated.

With a sample $X_1,\ldots,X_n \overset{i.i.d.}{\sim} P$, let $\bar{X}$ denote the sample mean, $\hat{\Sigma}_P$ denote the empirical covariance of the data, and $\hat{\Gamma}_P$ denote the empirical third-order central moments. Let $\sum_{j=1}^{d}\hat{\lambda}_j\hat{q}_j\otimes\hat{q}_j$ be a spectral decomposition of $\hat{\Sigma}_P$. Recall that the kth through dth eigenvalues of $\Sigma_{\mathbb{P}}$ must equal zero for every possible distribution in the model. Therefore, the variance of the data in each of the corresponding eigenvector directions is attributable to noise, so we suggest using $\hat{\sigma}^2 := \frac{1}{d-k+1}\sum_{j=k}^{d}\hat{\lambda}_j$ as an estimate of σ^2; a simpler alternative estimate is λ_k. As an approximate method of moments procedure, substitute $\bar{X}$ for $\bar{\mu}$, substitute

$$\hat{\Sigma}_{\mathbb{P}} = \sum_{j=1}^{k-1}(\hat{\lambda}_j-\hat{\sigma}^2)\hat{q}_j\otimes\hat{q}_j$$

for $\Sigma_{\mathbb{P}}$, and substitute $\hat{\Gamma}_P$ for $\Gamma_{\mathbb{P}}$ into the centered tensor algorithm. This isn't exactly a method of moments procedure, because there will typically be no model distribution whose moments match the empirical ones exactly.

Figure 1 visually demonstrates this method with six components in $\mathbb{R}^6$. Figure 2 demonstrates the convergence of our method in empirical total variation distance in the same setting. While the tensor method works well in this example, in some other simulations it not very robust to estimation error of the moments.

3.2 Gaussian Mixture with Polynomial Basis Expansion

If the number of components exceeds the number of variables, it can still be possible to use the tensor method after first expanding the points into a higher-dimensional space. Notice our application to Gaussian mixtures only used Normality to justify the estimation of σ^2; if σ^2 were known, we would only need to make sure that the average of the components' covariances can be estimated from the data. As a concrete example, let P be a mixture of up to five Gaussian components $\{\mu_1 = (\mu_{1,x},\mu_{1,y}),\ldots,\mu_k = (\mu_{k,x},\mu_{k,y})\}$ in $\mathbb{R}^2$, each with covariance $\sigma^2 I$ for a known σ^2. We will expand these points into a particular five-dimensional space in which the components are no longer all Gaussian but in which the average covariance can be estimated from the data.

Suppose without loss of generality that $\bar{\mu}$ equals the zero vector so that we do not have to carry the centering notation throughout our derivations. Additionally, let us suppose that the component points are rotated so that their correlation is zero. This is achieved by multiplying the centered points by the eigenvectors of $\Sigma_{\mathbb{P}}$ (that is, the principal components) which can be estimated using the relationship $\Sigma_{\mathbb{P}} = \Sigma_P - \sigma^2 I$. To avoid the extra notation, we'll assume that $\sum_j \mu_{j,x}\mu_{j,y} = 0$.

Consider the mapping

$$\phi : (x, y) \mapsto \begin{pmatrix} x, & y, & xy, & x^2, & y^2 \end{pmatrix}.$$

With $(X, Y) \sim P$ and with $J \sim \text{Unif}\{1, \ldots, k\}$ representing the choice of a mixture component, Section A.1 of the appendix calculates the conditional expectations of $\phi(X, Y)$ given J to be

$$(\mu_{J,x}, \quad \mu_{J,y}, \quad \mu_{J,x}\mu_{J,y}, \quad \mu_{J,x}^2 + \sigma^2, \quad \mu_{J,y}^2 + \sigma^2).$$

Let $\widetilde{P}$ denote the distribution of $\phi(X, Y)$, and let $\widetilde{\mathbb{P}}$ denote the uniform distribution on the *expanded component means*

$$\begin{aligned} \tilde{\mu}_1 &= (\mu_{1,x}, \quad \mu_{1,y}, \quad \mu_{1,x}\mu_{1,y}, \quad \mu_{1,x}^2 + \sigma^2, \quad \mu_{1,y}^2 + \sigma^2) \\ &\vdots \\ \tilde{\mu}_k &= (\mu_{k,x}, \quad \mu_{k,y}, \quad \mu_{k,x}\mu_{k,y}, \quad \mu_{k,x}^2 + \sigma^2, \quad \mu_{k,y}^2 + \sigma^2). \end{aligned}$$

By relating moments of these distributions, we will see how a tensor method enables us to estimate $\tilde{\mu}_1, \ldots, \tilde{\mu}_k$.

The overall expectations for $\widetilde{P}$ are the averages over J of the conditional expectations

$$\bar{\mu} := (0, \quad 0, \quad 0, \quad s_x^2 + \sigma^2, \quad s_y^2 + \sigma^2)$$

where $s_x^2 := \frac{1}{k}\sum_j \mu_{j,x}^2$ and $s_y^2 := \frac{1}{k}\sum_j \mu_{j,y}^2$. These are also the expectations of $\widetilde{\mathbb{P}}$. Therefore, with $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{i.i.d.}{\sim} P$, we can estimate $\bar{\mu}$ with the average of $\phi(X_1, Y_1), \ldots, \phi(X_n, Y_n)$. The second moments of $\mathbb{P}$ are

$$S_{\widetilde{\mathbb{P}}} = \bar{\mu}\bar{\mu}^T + \Sigma_{\widetilde{\mathbb{P}}}.$$

The covariance of $\widetilde{\mathbb{P}}$ coincides exactly with the covariance of the conditional expectation of $\widetilde{P}$ given J. Therefore, we can estimate $\Sigma_{\widetilde{\mathbb{P}}}$ by observing that $\Sigma_{\widetilde{P}}$ equals $\Sigma_{\widetilde{\mathbb{P}}}$ plus the expected conditional covariance of $\widetilde{P}$ given J, according to the law of total covariance. To this end, Section A.1 shows that the entries of $\phi(X, Y)$ are conditionally uncorrelated given J and have expected conditional variances

$$(\sigma^2, \quad \sigma^2, \quad s_x^2\sigma^2 + s_y^2\sigma^2 + \sigma^4, \quad 4s_x^2\sigma^2 + 2\sigma^4, \quad 4s_y^2\sigma^2 + 2\sigma^4).$$

The original coordinates' second moments s_x^2 and s_y^2 can be estimated from a sample $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{i.i.d.}{\sim} P$: the law of total variance justifies the estimators $\hat{s}_x^2 := \frac{1}{n}\sum_i X_i^2 - \sigma^2$ and $\hat{s}_y^2 := \frac{1}{n}\sum_i Y_i^2 - \sigma^2$.

Finally, rather than trying to deal with the third moments, it will be advantageous to directly estimate the third moments multiplied by the first coordinate vector

$$T_{\widetilde{\mathbb{P}}} \cdot e_1 = \frac{1}{k}\sum_j \mu_{j,x}\tilde{\mu}_j\tilde{\mu}_j^T.$$

The expectations of the entries of $X\phi(X, Y)\phi(X, Y)^T$ are described in Section A.2 so that the average of these matrices from the sample minus two *correction matrices* comprises an unbiased estimator

$$\mathbb{E}\left[\frac{1}{n}\sum_i X_i\phi(X_i, Y_i)\phi(X_i, Y_i)^T - \widehat{C}_1 - \widehat{C}_2\right] = T_{\widetilde{\mathbb{P}}} \cdot e_1$$

with

$$\widehat{C}_1 := \sigma^2 \begin{bmatrix} 0 & 0 & 0 & 2\hat{s}_x^2 & 0 \\ 0 & 0 & \hat{s}_x^2 & 0 & 0 \\ 0 & \hat{s}_x^2 & \hat{c}_{xxx} + \hat{c}_{xyy} & 2\hat{c}_{xxy} & 2\hat{c}_{xxy} \\ 2\hat{s}_x^2 & 0 & 2\hat{c}_{xxy} & 4\hat{c}_{xxx} & 0 \\ 0 & 0 & 2\hat{c}_{xxy} & 0 & 4\hat{c}_{xyy} \end{bmatrix}$$

and

$$\widehat{C}_2 := \sigma^2 \begin{bmatrix} 0 & 0 & 0 & 3\hat{s}_x^2 + 3\sigma^2 & \hat{s}_y^2 + \sigma^2 \\ 0 & 0 & \hat{s}_y^2 + \sigma^2 & 0 & 0 \\ 0 & \hat{s}_y^2 + \sigma^2 & 2\hat{c}_{xyy} & 3\hat{c}_{xxy} & \hat{c}_{yyy} \\ 3\hat{s}_x^2 + 3\sigma^2 & 0 & 3\hat{c}_{xxy} & 4\hat{c}_{xxx} & 2\hat{c}_{xyy} \\ \hat{s}_y^2 + \sigma^2 & 0 & \hat{c}_{yyy} & 2\hat{c}_{xyy} & 0 \end{bmatrix}$$

where

$$\hat{c}_{xxy} := \tfrac{1}{n} \sum_i X_i X_i Y_i$$

is an unbiased estimate of

$$\hat{c}_{xxy} := \tfrac{1}{k} \sum_j \mu_{j,x} \mu_{j,x} \mu_{j,z}$$

and likewise for other triplet combinations of coordinates.

Putting our estimated moments into the tensor method of Section 2.1 gives us estimates of the conditional expectations $\tilde{\mu}_1, \ldots, \tilde{\mu}_k$; the first two coordinates of those vectors are our estimates of $\mu_1, \ldots, \mu_k$, the component means in the original space.

This example readily generalizes to data with more than two variables. With d-dimensional data, one can follow our example to create an additional d quadratic variables along with $\binom{d}{2}$ product variables, one for every pair of coordinates.

With additional variables, new kinds of entries arise in the correction matrices. However, every possible type of entry appears in the case of $d = 5$ or larger. The reader can readily work out these terms or find them by inspecting a five-dimensional example in the code accompanying this paper. That code also includes the three-dimensional example shown in Figure 3. Figure 4 demonstrates the convergence of our method in empirical total variation distance in the same setting.

These simulations uses the expansion

$$\phi : (x, y, z) \mapsto \begin{pmatrix} x, & y, & z, & y^2, & z^2, & xy, & xz, & yz \end{pmatrix}.$$

We acknowledge that, based on this example and others, the method apparently requires a large amount of data to produce high quality estimates. Perhaps a better technique for modest sample sizes can be discovered by improving the moment estimation, by devising more robust tensor algorithms, or by making use of other basis transformations.

References

[Anandkumar et al., 2014] Anandkumar, A., Ge, R., Hsu, D., Kakade, S. M., and Telgarsky, M. (2014). Tensor Decompositions for Learning Latent Variable Models. *Journal of Machine Learning Research*, 15(1):2773–2832.

[Anandkumar et al., 2012] Anandkumar, A., Hsu, D., and Kakade, S. M. (2012). A method of moments for mixture models and hidden Markov models. In *Conference on Learning Theory*.

[Chang, 1996] Chang, J. T. (1996). Full Reconstruction of Markov Models on Evolutionary Trees: Identifiability and Consistency. *Mathematical Biosciences*, 137(1):51–73.

[Hsu and Kakade, 2013] Hsu, D. and Kakade, S. M. (2013). Learning mixtures of spherical Gaussians: moment methods and spectral decompositions. In *Innovations in Theoretical Computer Science*, pages 11–20. ACM.

A. Appendix

Lemma 1. *Let $u_1, \ldots, u_k$ be linearly independent vectors. If $\sum_{j=1}^{k} u_j \otimes u_j$ is the orthogonal projection operator onto their span, then $u_1, \ldots, u_k$ are orthonormal.*

Proof. The operator in question applied to u_i equals

$$\left[\sum_{j=1}^{k} u_j \otimes u_j\right] u_i = u_i + \sum_{j \neq i} \langle u_j, u_i \rangle u_j.$$

If it's the orthogonal projection operator, then the image is u_i which implies that $\langle u_j, u_i \rangle = 0$ for every $j \neq i$. Furthermore, this reveals that

$$\sum_{j=1}^{k} \|u_j\|^2 \frac{u_j}{\|u_j\|} \otimes \frac{u_j}{\|u_j\|}$$

is a spectral decomposition. Because any orthogonal projection's nonzero eigenvalues are all 1, we conclude that $u_1, \ldots, u_k$ must be unit vectors. □

A.1 Calculation of Conditional Covariance of Transformations

With $(X, Y) \sim P$ and with $J \sim \text{Unif}\{1, \ldots, k\}$ representing the choice of a mixture component, let us analyze the conditional covariances of the variables in $\phi(X, Y)$ given J. Letting $X = \mu_{J,x} + \sigma Z_x$ and $Y = \mu_{J,y} + \sigma Z_y$ with (Z_x, Z_y) each standard Normal and independent of each other and of J, we can derive the conditional expectations of the new variables

$$\begin{aligned}\mathbb{E}[XY|J] &= \mathbb{E}[(\mu_{J,x} + \sigma Z_x)(\mu_{J,y} + \sigma Z_y)|J] \\ &= \mu_{J,x}\mu_{J,y}\end{aligned}$$

$$\begin{aligned}\mathbb{E}[X^2|J] &= \mathbb{E}[(\mu_{J,x} + \sigma Z_x)^2|J] \\ &= \mu_{J,x}^2 + \sigma^2\end{aligned}$$

and analogously for Y^2. Notice that the mapping of the data in the fourth and fifth dimension is not exactly the same as the resulting mapping of the component means, as they differ by σ^2.

Now we are ready to work out the expected conditional covariances involving the new variables, starting with the variances.

$$\begin{aligned}\text{var}[XY|J] &= \mathbb{E}[(\sigma\mu_{J,x}Z_y + \sigma\mu_{J,y}Z_x + \sigma^2 Z_x Z_y)^2|J] \\ &= \sigma^2(\mu_{J,x}^2 + \mu_{J,y}^2 + \sigma^2)\end{aligned}$$

has average over J equal to $\sigma^2 s_x^2 + \sigma^2 s_y^2 + \sigma^4$.

$$\begin{aligned}\text{var}[X^2|J] &= \text{var}[(\mu_{J,x} + \sigma Z_x)^2|J] \\ &= 4\sigma^2\mu_{J,x}^2\text{var}[Z_x|J] + \sigma^4\text{var}[Z_x^2|J] \\ &= 4\sigma^2\mu_{J,x}^2 + 2\sigma^4\end{aligned}$$

has average over J equal to $4\sigma^2 s_x^2 + 4\sigma^4$. Analogously for Y^2.

$$\begin{aligned}\text{cov}[XY, X|J] &= \mathbb{E}[(\sigma\mu_{J,x}Z_y + \sigma\mu_{J,y}Z_x + \sigma^2 Z_x Z_y)\sigma Z_x|J] \\ &= \sigma^2\mu_{J,y}\end{aligned}$$

has average over J equal to zero because the component points are centered. Similarly for $\text{cov}[XY, Y|J]$.

$$\begin{aligned}\text{cov}[X^2, X|J] &= \mathbb{E}[(2\sigma\mu_{J,x}Z_x + \sigma^2 Z_x^2)\sigma Z_x|J] \\ &= 2\sigma^2\mu_{J,x}\end{aligned}$$

also has average over J equal to zero. Likewise for $\text{cov}[Y^2, Y|J]$. Finally,

$$\begin{aligned}\text{cov}[X^2, XY|J] &= \mathbb{E}[(2\sigma\mu_{J,x}Z_x + \sigma^2 Z_x^2)(\sigma\mu_{J,x}Z_y + \sigma\mu_{J,y}Z_x + \sigma^2 Z_x Z_y)|J] \\ &= 2\sigma^2\mu_{J,x}\mu_{J,y}\end{aligned}$$

and the same for $\text{cov}[Y^2, XY|J]$. Both have expectation over J equal to zero because we've assumed that the component points are centered and uncorrelated. All remaining pairs of different variables clearly have conditional covariance of zero due to the conditional independence of X and Y given J.

A.2 Expectations of Third Moment of Transformation Times First Coordinate Basis Vector

In this derivation, we will suppress the random vector argument, writing simply ϕ. To work out the expectation of $X\phi\phi^T$, we use of the representation $\phi = \tilde{\mu}_J + (\phi - \tilde{\mu}_J)$.

$$\begin{aligned}X\phi\phi^T &= \mu_{J,x}\phi\phi^T + \sigma Z_x\phi\phi^T \\ &= \mu_{J,x}\tilde{\mu}_J\tilde{\mu}_J + \mu_{J,x}[\tilde{\mu}_J(\phi - \tilde{\mu}_J)^T + (\phi - \tilde{\mu}_J)\tilde{\mu}_J^T] + \mu_{J,x}(\phi - \tilde{\mu}_J)(\phi - \tilde{\mu}_J)^T + \sigma Z_x\phi\phi^T\end{aligned}$$

The first term has expectation equal to the true $T_{\widetilde{P}} \cdot e_1$ which we want to estimate. To this end, we can subtract from $\frac{1}{n}\sum_i X_i\phi_i\phi_i^T$ estimates of the expectations of the three additional matrices; we will label these expectations C_0, C_1, and C_2 respectively.

The expectation of each entry of each matrix can be derived. Here we will merely demonstrate the process by working out the $(3, 4)$ entry of each of the three matrices. Starting with C_2, the $(3, 4)$ entry is

$$\begin{aligned}&\mathbb{E}\,\sigma Z_x[XY][X^2] \\ &= \mathbb{E}\sigma Z_x[(\mu_{J,x} + \sigma Z_x)(\mu_{J,y} + \sigma Z_y)][(\mu_{J,x} + \sigma Z_x)^2] \\ &= \mathbb{E}\,\sigma Z_x[\mu_{J,x}\mu_{J,y} + \sigma\mu_{J,x}Z_y + \sigma\mu_{J,y}Z_x + \sigma^2 Z_x Z_y][\mu_{J,x}^2 + 2\sigma\mu_{J,x}Z_x + \sigma^2 Z_x^2] \\ &= \mathbb{E}\,\sigma Z_x\mu_{J,x}\mu_{J,y}2\sigma\mu_{J,x}Z_x + \mathbb{E}\sigma Z_x\sigma\mu_{J,y}Z_x\mu_{J,x}^2 + \text{terms with expectation } 0 \\ &= 3\sigma^2 c_{xxy}.\end{aligned}$$

The $(3, 4)$ entry of C_1 is

$$\begin{aligned}&\mathbb{E}\,\mu_{J,x}[XY - \mu_{J,x}\mu_{J,y}][X^2 - (\mu_{J,x}^2 + \sigma^2)] \\ &= \mathbb{E}\mu_{J,x}[\sigma\mu_{J,x}Z_y + \sigma\mu_{J,y}Z_x + \sigma^2 Z_x Z_y][2\sigma\mu_{J,x}Z_x + \sigma^2 Z_x^2 - \sigma^2] \\ &= \mathbb{E}\,\mu_{J,x}\sigma\mu_{J,y}Z_x 2\sigma\mu_{J,x}Z_x + \text{terms with expectation } 0 \\ &= 2\sigma^2 c_{xxy}.\end{aligned}$$

Finally C_0 has as its $(3, 4)$ entry

$$\begin{aligned}&\mathbb{E}\,\mu_{J,x}[(\mu_{J,x}\mu_{J,y})(X^2 - (\mu_{J,x}^2 + \sigma^2)) + (\mu_{J,x}^2 + \sigma^2)(XY - \mu_{J,x}\mu_{J,y})] \\ &= \mathbb{E}\,\mu_{J,x}[(\mu_{J,x}\mu_{J,y})(2\sigma\mu_{J,x}Z_x + \sigma^2 Z_x^2 - \sigma^2) + (\mu_{J,x}^2 + \sigma^2)(\sigma\mu_{J,x}Z_y + \sigma\mu_{J,y}Z_x + \sigma^2 Z_x Z_y)]\end{aligned}$$

which can be shown to equal zero. Conveniently, every entry of C_0 turns out to be zero, so it does not require a corresponding correction matrix.

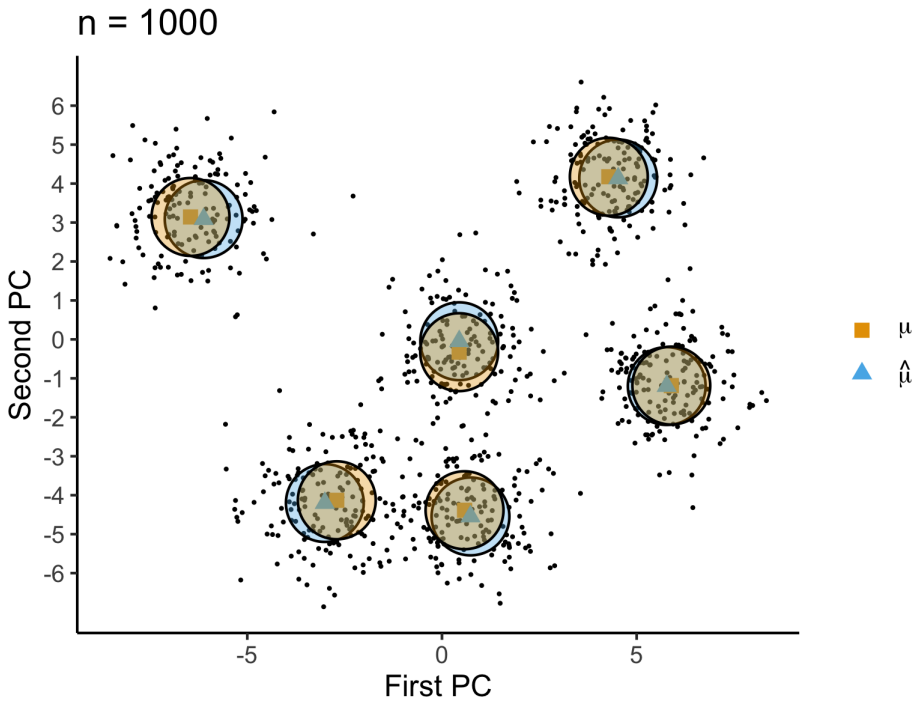

Figure 1: True component mean (orange) compared to their estimated values using our method (orange) at sample size $n = 10^3$. Circle radii indicate $\sigma = 1$ (blue) compared to $\hat{\sigma} = 0.99$ (orange). The $k = 6$ component means were drawn from the spherical Gaussian distribution on $\mathbb{R}^6$ centered at the origin and with covariance 9 times the identity matrix. The component distributions themselves are Gaussian with covariance equal to the identity matrix.

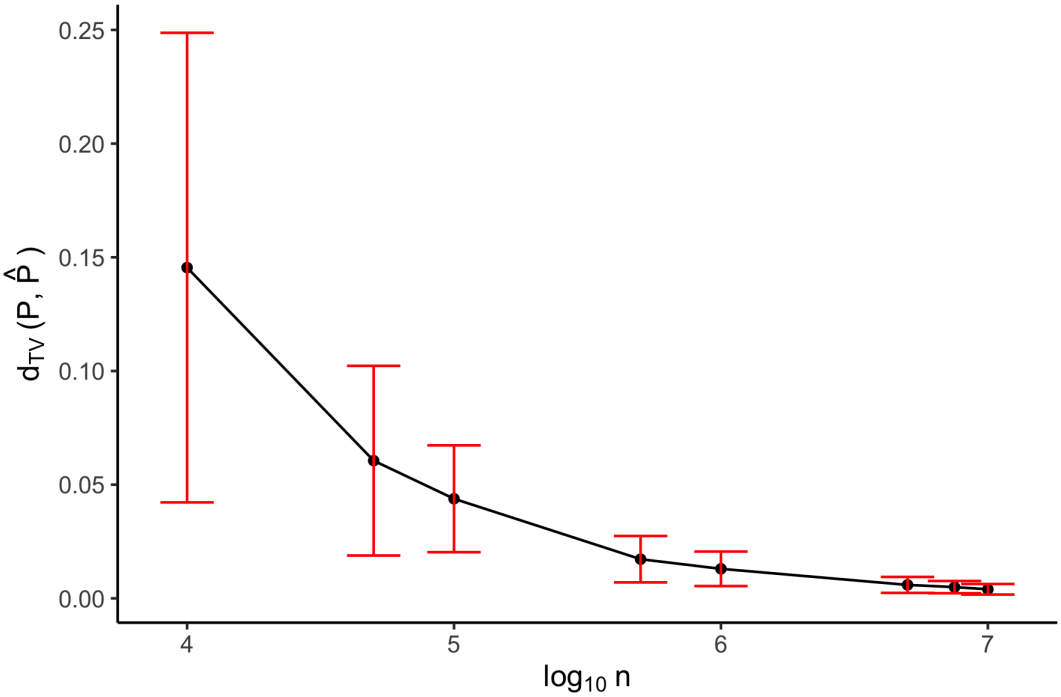

Figure 2: Empirical Total Variation Distance between true distribution (P) and estimated distribution ($\hat{P}$) v.s. sample size (n). The $k = 6$ component means were drawn from the spherical Gaussian distribution on $\mathbb{R}^6$ centered at the origin and with covariance 9 times the identity matrix. The component distributions themselves are Gaussian with covariance equal to the identity matrix. The components are Gaussian with covariance equal to the identity matrix.

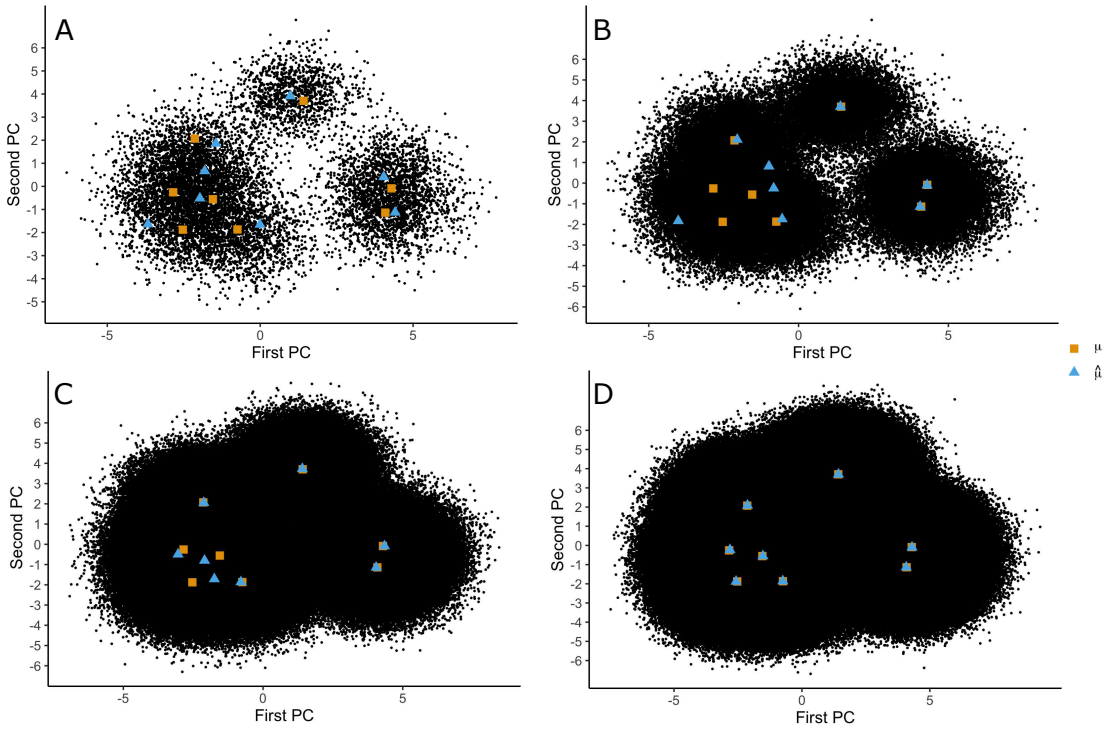

Figure 3: True component mean (orange) compared to their estimated values using our method (orange) at sample sizes (A) $n = 10^4$ (B) $n = 10^5$ (C) $n = 10^6$ (D) $n = 10^7$. The $k = 8$ component means were drawn from the spherical Gaussian distribution on $\mathbb{R}^3$ centered at the origin and with covariance 9 times the identity matrix. The component distributions themselves are Gaussian with covariance equal to the identity matrix.

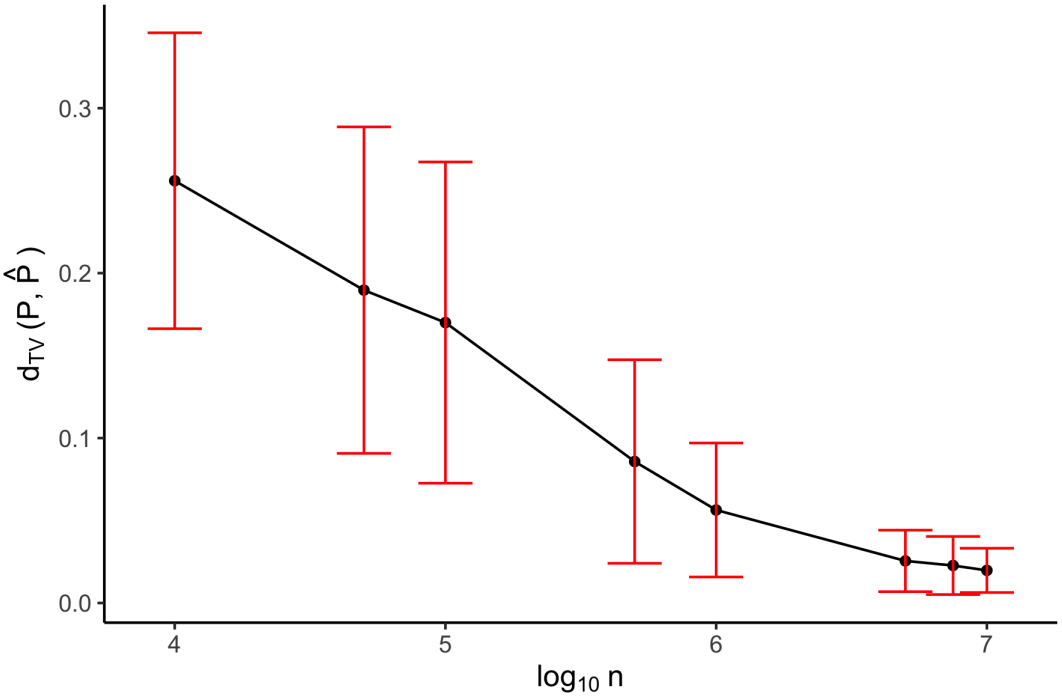

Figure 4: Empirical Total Variation Distance between true distribution (P) and estimated distribution ($\hat{P}$). The $k = 8$ component means were drawn from the spherical Gaussian distribution on $\mathbb{R}^3$ centered at the origin and with covariance 9 times the identity matrix. The component distributions themselves are Gaussian with covariance equal to the identity matrix.